



universität  
wien

# DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

“CLIL with a capital I: Using cognitive discourse functions to integrate content and language learning in CLIL history education”

verfasst von / submitted by

Mag.phil. Silvia Bauer-Marschallinger

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of  
Doktorin der Philosophie (Dr.phil.)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt /  
degree programme code as it appears on the student  
record sheet:

UA 792 343

Dissertationsgebiet lt. Studienblatt /  
field of study as it appears on the student record sheet:

English and American Studies

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Christiane Dalton-Puffer



## Acknowledgements

Fitting the topic of the present study, I will start this section by taking a historical perspective. I'd say that my journey started when I was working on my MA thesis under the supervision of Christiane Dalton-Puffer, who was an excellent and dedicated supervisor then and who has been an even more helpful and encouraging mentor now during my PhD studies. I'd like to express my heartfelt thanks to her. Throughout the project, Christiane has been nothing but supportive, understanding, and patient, always offering helpful feedback and advice – not just for the thesis but for navigating academic life more generally.

It was also Christiane who suggested applying for a uni:docs grant, which brings me to my next point. I would like to thank the uni:docs programme for funding this project and making it possible for me to spend three wonderful years as uni:docs fellow at the Department of English and American Studies at the University of Vienna. During my time at the English Department, I had the chance to share my research journey with fellow pre-docs, whom I could always count on for research-related and emotional support. I'd like to especially thank my office buddies Sarah, Marlene, Elli, Christina, Alex, Vanja, & Kathi. Thank you for listening to my challenges, answering questions, and providing advice whenever needed, but also for all the fun experiences we've shared. Furthermore, I'd like to express my gratitude to the EduLing research group at the English Department. Your constructive feedback and kind encouragement were of great help throughout my project.

Elsewhere, I'd like to thank Lena Heine for providing feedback on the initial drafts of my pilot teaching materials and Ana Llinares and Tarja Nikula for sharing their thoughts about coding with the CDF construct, leading me to insights useful for my own coding. I would also like to express my gratitude to Greta Wurm, who has transcribed my data diligently yet swiftly, and to Tatjana Bacovsky for reading a huge part of this thesis, finding typos and offering helpful comments.

Next, I'd like to thank the two schools I've been cooperating with. A special thanks goes out to the teachers for allowing me into their classroom and taking the time to create and improve the materials designed in the course of this study. Their expertise but also their readiness and continued interest have been invaluable. Moreover, I'd like to thank their students for agreeing to take part and sharing their perspectives so openly.

Finally, I'd like to thank those people who have been there for me on a more general level. I'd like to thank my friends, especially the "Leischen-collective" and the ESN girls. Thanks for listening to me talk and sometimes rant about work-related things and, even more importantly, for all the great moments we've shared that distracted me from feelings of stress and worry. I would also like to express my gratitude to my family. I'd like to thank my siblings, who have taken similar paths, for giving me academia-related advice but also for providing welcome distractions from work. A heartfelt thanks goes out to my parents, who have encouraged me to be curious from a young age onwards and who have been supportive throughout my whole education. For this and

more, I'll be forever grateful. Last, I'd like to thank Alex for being such a wonderful partner. From being my "local" IT guy and office buddy in our home office, to cooking tasty meals, making sure I'm taking enough breaks, and encouraging me whenever I needed it - you have always been there for me.

Now, looking back at my "PhD journey", all I can say is that I am immensely grateful for having such wonderful people in my life.

## Abstract

Being an educational approach that was primarily introduced to innovate language instruction, it is not surprising that Content and Language Integrated Learning (CLIL) has mostly been researched from the perspective of applied linguistics. Concerns relating to subject learning, in contrast, have only recently started to gain attention. With subject learning taking on a greater role in CLIL research, the content-and-language-integrative nature of this educational approach has become one of the central themes in the field. Conceptually, several propositions have been made concerning the integration of content and language learning, many of which aligning with systemic functional linguistics and/ or sociocultural theory. While these theoretical approaches have yielded interesting insights into the integration of subject and language learning, they do not translate into classroom practice easily. One notion allowing conceptual integration while appearing to be tangible for practitioners is the construct of cognitive discourse functions (CDFs; Dalton-Puffer, 2013). Being both anchored in linguistics and education, CDFs are assumed to be the generic linguistic manifestation of cognitive processes essential to learning and teaching. In the field of history education, too, CDFs have been shown to be tightly linked to history skills, both conceptually and empirically. Thus far, however, this construct has not been operationalized for pedagogical use, and generally more research is needed concerning the nexus of content-and-language-integrative learning, pedagogical practice, and didactic materials, also considering that CLIL teachers urgently lack integrative material as well as conceptual understanding in this respect.

To address this gap, this PhD project is set in a framework of design-based research (DBR), which has been heralded as a transdisciplinary methodological approach able to reconcile theory- and practice-related concerns by being dual-focused. As such, this thesis aims to (1) further illuminate the theoretical underpinnings of the integration of content and language learning and (2) to develop practice-oriented tools and materials for upper secondary CLIL history education. With these aims in mind, I closely collaborated with teachers in order to systematically develop CDF-based history materials. First, the needs of participants were determined using individual interviews with teachers, focus group interviews with students, and written competency-based task for the learners, which informed the intervention we designed. Then, the teacher implemented these materials in their own class. Finally, the process and the products were evaluated from the learners' and the teacher's perspective as well as via written learner tasks once again. Based on these findings, our approach and the materials were advanced and fine-tuned over three such research cycles in two contexts.

The findings of this study have shown that CDFs present an ecologically valid and effective approach to integrate content and language learning in upper secondary CLIL history education. Yet, for these materials to be accepted and to take effect, several conditions need to be met: First of all, competency-based tasks need to be engaging, interactive, and scaffolded in small steps, and the links between the linguistic support and the subject discipline need to be made explicit. Moreover, such scaffolding should not only consider linguistic forms and functions but also

subject-specific concepts and notions important in the discipline. Additionally, in the course of the project, the importance of differentiated instruction crystallized. These aspects were crucial for the participants' acceptance of the new approach, which also seemed to be reflected in the learners' performance. Initially, both groups involved in the main study struggled with demonstrating subject-specific skills in English in various domains, such as appropriately justifying claims, signalling communicative intentions, or linking ideas. In the case of group A, who received two treatments, ratings improved significantly both in terms of academic language skills and history competences, with the bigger leap in performance in their second round. In contrast, the scores in group B, who received one treatment, increased only moderately (but statistically significantly) in the linguistic domain, while content results remained steady. Finally, this thesis has also demonstrated that the CDF construct is a useful and manageable tool for research. Yet, to ensure reliable coding, further specifications for different subjects may be needed, which this thesis intends to provide for the subject history.

# Table of Contents

<b>Acknowledgements</b> .....	<b>iii</b>
<b>Abstract</b> .....	<b>v</b>
<b>List of Figures</b> .....	<b>xi</b>
<b>List of Tables</b> .....	<b>xii</b>
<b>List of abbreviations</b> .....	<b>xiii</b>
<b>List of statistical symbols and abbreviations</b> .....	<b>xiv</b>
<b>1. Introduction</b> .....	<b>1</b>
1.1 Positioning of the thesis .....	1
1.2 Rationale and aims .....	2
1.3 Outline.....	3
1.4 Key terms.....	4
<b>2. Content and Language Integrated Learning (CLIL)</b> .....	<b>6</b>
2.1 Origins and labels.....	6
2.2 CLIL implementation in Austria .....	7
2.3 Recent developments in CLIL research.....	9
2.3.1 Looking back: research on Canadian and American bilingual education .....	9
2.3.2 Outcome-oriented CLIL studies.....	10
2.3.3 Participant perspectives .....	15
2.3.3.1 Affective factors .....	16
2.3.3.2 Beliefs and perceptions .....	18
2.3.3.3 Participants' voices in didactic design.....	20
2.3.4 Materials and didactic design in CLIL.....	21
2.3.5 CLIL and diversity.....	26
2.3.6 Linking the strands .....	29
<b>3. Conceptualizing the integration of content and language</b> .....	<b>32</b>
3.1 Sociocultural theory.....	33
3.2 Systemic functional linguistics.....	37
3.2.1 Key concepts and overview.....	38
3.2.2 Subject literacy: the case of history .....	39
3.2.2.1 History genres .....	40
3.2.2.2 Features of historical language.....	41
3.2.3 SFL-based approaches to integrate content and language learning in bilingual history education .....	44
3.3 From the 4Cs framework to a pluriliteracies approach.....	49
3.4 A construct of cognitive discourse functions (CDFs) .....	51
3.4.1 Rationale and aims of the CDF construct .....	51
3.4.2 Theoretical background.....	53
3.4.3 Features of the CDF construct (Dalton-Puffer, 2013).....	58
3.4.4 The 7 CDF types.....	59
3.4.4.1 CATEGORIZE/ CLASSIFY.....	59
3.4.4.2 DEFINE .....	60
3.4.4.3 DESCRIBE .....	60

3.4.4.4	EVALUATE.....	61
3.4.4.5	EXPLAIN .....	62
3.4.4.6	EXPLORE .....	62
3.4.4.7	REPORT .....	63
3.4.5	Empirical validation of the construct.....	64
3.4.6	Operationalizing the CDF construct .....	70
<b>4.</b>	<b>History education.....</b>	<b>73</b>
4.1	Central notions and themes in history didactics.....	73
4.1.1	Historical competence .....	73
4.1.2	Historical thinking.....	74
4.1.3	Historical consciousness .....	76
4.1.4	Historical literacy.....	77
4.1.5	Historical reasoning.....	77
4.2	The FUER Model .....	78
4.2.1	Questioning competence .....	79
4.2.2	Methodological competence .....	80
4.2.3	Orientation competence .....	81
4.2.4	Factual competence .....	82
4.2.5	Grading Matrix and performative verbs .....	83
4.2.6	Empirical validation and criticism of the FUER model .....	85
4.3	History education in Austria .....	86
4.3.1	The Austrian school system .....	86
4.3.2	The student body of vocational upper secondary education .....	88
4.3.3	History education in secondary colleges of business administration (HAK).....	88
<b>5.</b>	<b>Methodology .....</b>	<b>91</b>
5.1	Design-based research (DBR).....	91
5.1.1	Typical characteristics of DBR and terminology.....	91
5.1.2	Methods and research design in DBR .....	93
5.1.3	Practical outcomes and theorizing in DBR .....	95
5.2	Context of the present study .....	95
5.2.1	School A.....	96
5.2.2	School B.....	96
5.3	Research design .....	97
5.3.1	Participants .....	99
5.3.1.1	Teachers.....	99
5.3.1.2	Students.....	100
5.3.2	Ethical considerations .....	102
5.3.3	Key steps of a research cycle .....	103
5.3.3.1	Needs analysis.....	103
5.3.3.2	Intervention design.....	104
5.3.3.3	Implementation .....	104
5.3.3.4	Evaluation.....	104
5.3.4	Organisation of cycles.....	105
5.4	Data collection and instruments .....	106
5.4.1	Interviews .....	107
5.4.1.1	Semi-structured interviews with teachers: needs analysis .....	109



5.4.1.2	Focus group interviews with students: needs analysis .....	109
5.4.1.3	Retrospective focus group interview with students: evaluation.....	110
5.4.1.4	Retrospective interviews with teachers: evaluation.....	111
5.4.2	Competency-based written tasks .....	111
5.4.3	Intervention design-sessions .....	114
5.4.4	Observations (online/ offline) .....	114
5.5	Data analysis.....	115
5.5.1	Transcription .....	116
5.5.2	Qualitative content analysis .....	116
5.5.3	Coding and rating of written tasks.....	120
5.5.3.1	The CDF-based coding and linguistic rating .....	121
5.5.3.2	Subject-specific rating.....	123
5.5.3.3	Statistical procedures.....	127
5.5.4	Field notes and lesson transcripts .....	129
<b>6.</b>	<b>Pilot study .....</b>	<b>130</b>
<b>7.</b>	<b>Analysis.....</b>	<b>133</b>
7.1	Needs analysis .....	133
7.1.1	Semi-structured interviews with the teachers.....	134
7.1.1.1	Pedagogical practices and beliefs related to language .....	134
7.1.1.2	Materials: selection, adaptation, and creation .....	138
7.1.1.3	Learner needs: subject and language learning .....	140
7.1.1.4	Summary and implications for the design .....	143
7.1.2	Focus group interviews with the students .....	144
7.1.2.1	Views on BE and previous experiences: benefits and problems .....	144
7.1.2.2	Pedagogical practices in the bilingual classroom related to language .....	148
7.1.2.3	Bilingual materials: status quo and wishes for the future .....	150
7.1.2.4	Learner needs: subject and language learning .....	151
7.1.2.5	Summary and implications for the design .....	153
7.1.3	Initial competency-based written tasks .....	155
7.1.3.1	History-based rating results .....	156
7.1.3.2	Linguistic rating results and their connections to content results .....	160
7.1.3.3	Summary and implications for the design .....	165
7.2	Designing the intervention .....	167
7.2.1	Modus operandi of design sessions .....	168
7.2.2	Unit I: absolutism and mercantilism (cycle 1 & 2).....	170
7.2.2.1	Design process of unit I for school A (cycle 1) .....	170
7.2.2.2	Adaption of unit I for school B (cycle 2) .....	175
7.2.2.3	Final lesson plan of unit I.....	177
7.2.2.4	Final version of worksheets of unit I.....	179
7.2.3	Unit II: the Industrial Revolution (cycle 3).....	188
7.2.3.1	Design process of unit II.....	188
7.2.3.2	Final lesson plan of unit II.....	191
7.2.3.3	Final version of worksheets of unit II .....	193
7.3	Implementation phase .....	204
7.3.1	Cycle 1 .....	204
7.3.2	Cycle 2 .....	207

7.3.3	Cycle 3 .....	209
7.4	Evaluation of interventions .....	211
7.4.1	Cycle 1: absolutism and mercantilism (school A) .....	213
7.4.1.1	Retrospective interviews with students and teacher .....	213
7.4.1.2	Post-intervention written tasks .....	222
7.4.2	Cycle 2: absolutism and mercantilism (school B) .....	230
7.4.2.1	Retrospective interviews with students and teacher .....	230
7.4.2.2	Post-intervention written tasks .....	241
7.4.3	Cycle 3: the Industrial Revolution (school A).....	251
7.4.3.1	Retrospective interviews with students and teacher .....	251
7.4.3.2	Post-intervention written tasks .....	262
<b>8.</b>	<b>Discussion .....</b>	<b>273</b>
8.1	Answers to the research questions, including design principles.....	273
8.1.1	RQ1: learner needs and features of materials.....	273
8.1.2	RQ2: reactions by learners and teachers.....	286
8.1.2.1	Perceptions of students.....	286
8.1.2.2	Perceptions of teachers.....	289
8.1.3	RQ3: the effect on learner language and content learning.....	290
8.1.3.1	Overall results: content and language .....	290
8.1.3.2	History outcomes.....	291
8.1.3.3	Language outcomes.....	293
8.1.3.4	Achievement groups.....	295
8.2	Conceptual discussion: the CDF construct revisited .....	295
8.3	Methodological insights .....	302
<b>9.</b>	<b>Conclusion.....</b>	<b>305</b>
9.1	Summary .....	305
9.2	Key findings and implications .....	308
9.2.1	Design principles.....	308
9.2.2	The participants' reactions to the intervention .....	311
9.2.3	The effect of CDF-oriented teaching on the learners' historical competences and academic language skills .....	312
9.3	Significance, limitations, and outlook .....	314
	<b>References.....</b>	<b>320</b>
	<b>Deutsche Zusammenfassung .....</b>	<b>358</b>
	<b>Digital appendix .....</b>	<b>360</b>

## List of Figures

Figure 1. 4Cs model (Coyle et al., 2010, p. 41) .....	49
Figure 2. Mapping pluriliteracies development (Meyer et al., 2015, p. 49).....	50
Figure 3. Representation of the FUER model based on Körber et al. (2007) and slightly modified from Dalton-Puffer and Bauer-Marschallinger (2019, p. 37).....	80
Figure 4. A schematic representation of the Austrian school system provided by the Austrian Federal Ministry for Education (2021).....	87
Figure 5. DBR as a cyclical research process based on Fraefel (2014) and McKenney & Reeves (2012) .....	94
Figure 6. Overview of the research design and data collection.....	98
Figure 7. Code trees: needs analysis teacher interview .....	119
Figure 8. Linguistic rubric.....	122
Figure 9. Notes sheet for the linguistic rating.....	123
Figure 10. Level rating rubric.....	124
Figure 11. Competence rubric targeting deconstruction competence .....	125
Figure 12. Competence rubric focusing on orientation competence.....	126
Figure 13. Spreadsheet to document history ratings .....	126
Figure 14. Worksheet Louis XIV, source A, task 3 .....	170
Figure 15. Revision: worksheet Louis XIV, source A, task a and b (later relabelled as task 1 and 2) .....	173
Figure 16. Revision: worksheet Louis XIV, source B, task b (later relabelled as task 2) .....	175
Figure 17. Revised instructions: worksheet Louis XIV .....	175
Figure 18. IR – part 3: impulse question and task c .....	189
Figure 19. IR – part 4b: CATEGORIZE activity .....	190
Figure 20. IR – part 3c: fast-track activity.....	190
Figure 21. Code co-occurrence model: experience of students and teacher, group A, cycle 1.....	216
Figure 22. Code co-occurrence model: evaluation by the students of group A, cycle 1 .....	216
Figure 23. Code co-occurrence model: evaluation by T <sub>A</sub> , cycle 1 .....	217
Figure 24. Comparison T1 vs. T2: boxplots, group A.....	222
Figure 25. Linguistic rating: T1 vs. T2, group A, cycle 1 .....	226
Figure 26. Code co-occurrence model: evaluation by the students of group B.....	233
Figure 27. Code co-occurrence model: evaluation by T <sub>B2</sub> .....	234
Figure 28. Comparison T1 vs. T2: boxplots, group B.....	241
Figure 29. History-based rating: T1 vs. T2, group B, cycle 2.....	244
Figure 30. Code co-occurrence model: experience of students and teacher, group A, cycle 3.....	251
Figure 31. Code co-occurrence model: cycle 3 evaluation by the students of group A .....	255
Figure 32. Code co-occurrence model: cycle 3 evaluation by T <sub>A</sub> .....	256
Figure 33. ETS12's markings on the handout .....	258
Figure 34. Comparison T1, T2, & T3: boxplots, group A.....	262
Figure 35. Development of written task results: group A, divided into achievement groups based on T1 performance.....	263
Figure 36. Development of written task results: group A, divided into achievement groups based on previous grades.....	264
Figure 37. Development of history results, group A .....	265
Figure 38. Overlaps of CDF types in the written tasks with a focus on REPORT: T1 (left) vs. T3 (right) .....	266
Figure 39. Code co-occurrence model of all CDF codings of all written task performances .....	301

## List of Tables

Table 1. Main genres in history (based on Christie & Derewianka, 2008; Coffin, 2006; Llinares & Pascual Peña, 2015).....	40
Table 2. Sample descriptors for speaking in history/ civics and mathematics (Moe et al., 2015, pp. 69–70) .....	56
Table 3. The CDF construct (Dalton-Puffer & Bauer-Marschallinger, 2019, p. 35).....	58
Table 4. Characteristics of design-based research according to Wang and Hannafin (2005, p. 8) .....	92
Table 5. Research questions and methods .....	106
Table 6. Research questions and methods of data collection and analysis.....	115
Table 7. Overview of results: T1 .....	156
Table 8. Average history-based ratings .....	157
Table 9. Average linguistic ratings: T1 .....	160
Table 10. Individual results: group A, cycle 1 .....	223
Table 11. History-based rating: T1 vs. T2, group A, cycle 1.....	224
Table 12. Individual results, group B, cycle 2.....	242
Table 13. Differences T2-T1, divided into achievement groups based on their T1 performance .....	242
Table 14. Differences T2-T1, divided into achievement groups based on English and history grades.....	243
Table 15. Linguistic rating: T1 vs.T2, group B, cycle 2.....	246
Table 16. History-based rating: differences between T1, T2, & T3, group A .....	265
Table 17. Linguistic rating: differences T1, T2, & T3, group A .....	267
Table 18. Revision of Dalton-Puffer’s CDF construct for the subject history.....	298

## List of abbreviations

BE	bilingual education
CA	CATEGORIZE (CDF type)
CBI	content-based instruction
CDF	cognitive discourse function
CLIL	content and language integrated learning
DBR	design-based research
DF	DEFINE (CDF type)
DS	DESCRIBE (CDF type)
EA	EXPLAIN (CDF type)
EFL	English as a foreign language
EO	EXPLORE (CDF type)
EV	EVALUATE (CDF type)
FC	factual competence
FL	foreign language
FUER	Für die Förderung und Entwicklung von reflektiertem Geschichtsbewusstsein
IR	Industrial Revolution
L1	first language
L2	second language (including all languages learned after the L1)
MC	methodological competence
OC	orientation competence
QCA	qualitative content analysis
R2L	reading to learn
RE	REPORT (CDF type)
SCT	sociocultural theory
T1	pre-intervention task
T2	post-intervention task
T3	second post-intervention task
TA	teacher A
TB1	teacher B1
TB2	teacher B2
TLC	teaching/ learning cycle
WS	worksheet
ZPD	zone of proximal development

## List of statistical symbols and abbreviations

$\alpha$	significance level
$\chi^2$	chi square test value
$d$	Cohen's $d$
$df$	degrees of freedom
$F$	$F$ -ratio
$\eta^2_g$	generalized eta squared
$ICC$	intraclass correlation
$\tau_b$	Kendall's tau b
$Max$	maximum
$M$	mean
$Mdn$	median
$Min$	minimum
$r_p$	Pearson's correlation coefficient
$p$	probability
$R$	range
$n$	size of subsample
$N$	size of complete data set
$SD$	standard deviation
$t$	t-statistic ( $t$ -test)
$T$	T-statistic (Wilcoxon signed-rank test)
$r_w$	Wilcoxon effect size
$Z$	z-score

# 1. Introduction

## 1.1 Positioning of the thesis

*Every teacher is a language teacher.* This often quoted notion goes back to the 1970s when policymakers and educators acknowledged the educational role of language across the curriculum (van der Walt & Ruiters, 2012, p. 85). This idea that language, literacy, and content learning are interconnected has, unsurprisingly, struck a chord with bilingual programmes, which are expanding worldwide. In Europe, the predominant model is called CLIL – *content and language integrated learning* – which, according to the label, stresses the importance of fusing the learning of content and language. Historically, however, the objective of this approach was to innovate and improve language learning (Coyle et al., 2010; European Commission, 1995), and consequently CLIL research has for a long time mostly focused on linguistic aspects (Dafouz et al., 2014; Dalton-Puffer, 2018; San Isidro, 2019). While content learning has started to receive more attention in recent years (Dalton-Puffer, 2018; Fernández-Sanjurjo et al., 2017; San Isidro, 2019), many of these studies conceptualize content learning as declarative knowledge, i.e., knowing facts and figures (e.g., Dallinger et al., 2016 or Gablasova, 2014). This seems to be in contrast to modern curricula, which, by and large, define learning aims via subject-specific competences rather than topics and areas of knowledge (Gautschi, 2015; Priestley & Biesta, 2017). Taking a look at CLIL practice, it appears that CLIL has been determined more strongly by content-related considerations rather than language didactics (Dalton-Puffer et al., 2010; Nikula et al., 2016). However, the way CLIL is practised is very diverse and dependent on the local context (Eurydice, 2017). This fact is reflected in many definitions of CLIL which stress that “CLIL” is an umbrella term (Cenoz et al., 2014; Dalton-Puffer et al., 2014) for any educational approach that is “dual-focused” (Coyle et al., 2010, p. 1) and “where a language other than the students’ mother tongue is used as medium of instruction” (Dalton-Puffer, 2007, p. 1).

These definitions, however, imply that *language* and *content* are two separate domains, which seems contrary to its label. Recently, however, researchers have started to pay more attention to how these two perspectives can be integrated. For some, this integration even is “the way forward in [...] CLIL for the rest of the twenty-first century” (Ruiz de Zarobe & Cenoz, 2015, p. 90). Such discussions tie in with other current developments in CLIL research. While early studies into the effectiveness of CLIL tended to report positive results for different domains of language learning and neutral results for content learning, more recent studies often report less optimistic findings, which overall raises questions about CLIL’s efficacy and whether CLIL lives up to its potential (Dalton-Puffer, 2008; Graham et al., 2018; Pérez Cañado, 2016b; San Isidro, 2019). From the current perspective, it appears that genuine integration of language and content learning could be one of our best bets to fulfil the great promise of CLIL, namely a more effective pedagogy of bilingual education that leads to satisfactory learning outcomes both for the subject and the

foreign language (see, e.g., Meyer et al., 2015; Morton, 2020; Nikula et al., 2016; Pavón Vázquez, 2018).

The importance of theorizing this integration seems undisputed and several theoretical models have been suggested for its conceptualization, most of which are situated in a framework of systemic functional linguistics and/or sociocultural theory (e.g., Lo & Jeong, 2018; Meyer & Coyle, 2017; J. Moore et al., 2018; Rose & Martin, 2012; Tedick & Lyster, 2019). However, these frequently do not translate easily into classroom practice for various reasons. One factor, for sure, seems to be that these pedagogical innovations are often deeply rooted in linguistics, which runs the risk of creating solutions that (content) CLIL teachers might find too abstract or even transgressive (Dalton-Puffer & Bauer-Marschallinger, 2019; Schall-Leckrone & McQuillan, 2012). Moreover, teachers lack suitable materials and often struggle with creating or adapting their own materials due to a lack of time resources but also limited training and expertise in designing materials that interlink the learning of content and a foreign language (see, e.g., Gruber et al., 2020; Hahn, 2019; Morton, 2013; Pérez Cañado, 2016a; Skinnari, 2020).

What is needed, then, is research which not only operationalizes scientific insights into content and language integration for pedagogical use but research that tries to transcend the boundaries of the disciplines, those being educational linguistics and subject-specific didactics. One notion that intends to be such a “zone of convergence” (Dalton-Puffer, 2013, p. 216) is the concept of *cognitive discourse functions* (CDFs). CDFs are linguistic patterns that are routinely used when engaging with content, e.g., when we explain or evaluate (Dalton-Puffer, 2013, 2016). As such, they resemble performative verbs used in curricula and competency-based testing, and consequently CDFs are believed to have face validity with learners and teachers (Dalton-Puffer, 2013). Having reviewed a number of frameworks and taxonomies of academic and subject-specific language functions, Dalton-Puffer (2013) suggested a construct of seven central CDF types. While this construct has been empirically validated in a number of small-scale studies (e.g., Breeze & Dafouz, 2017; Dalton-Puffer et al., 2018; Doiz & Lasagabaster, 2021; Lorenzo, 2017; Nashaat-Sobhy & Llinares, 2020), research into the pedagogical operationalization of the model is still in its infancy. The following thesis now attempts to address this issue for the subject history at upper secondary level.

## **1.2 Rationale and aims**

In order to create a viable pedagogical design while also driving forward the theoretical underpinnings of the integration of language and content learning, I chose design-based research (DBR) as the methodological framework for this study. Being “heralded as a practical research methodology that could effectively bridge the chasm between research and practice in formal education” (T. Anderson & Shattuck, 2012, p. 16), DBR usually pursues two aims, namely solving educational problems via systematically developing viable didactic tools and driving forward its underlying theory. These are also the two aims this thesis intends to accomplish. To be more precise, this study strives to identify features of effective content-and-language-integrative



materials, which enable secondary CLIL learners of history to better verbalize their cognitive processes, i.e., their CDF use, and to develop blueprint materials accordingly. Typical for DBR, the design of this pedagogical intervention is characterized by close collaboration with the practitioners involved in this study in order to create practical and robust pedagogical products (e.g., McKenney & Reeves, 2012, 2014; van den Akker & Nieveen, 2016). Three former teacher colleagues at two different upper secondary vocational schools with a focus on business were chosen for this collaboration. In terms of research process, this study is organized in five consecutive research cycles, two functioning as pilot and three as main cycles. These cycles consisted of four stages: First, the local context and the needs of the participants were explored via interviews with the teacher and a focus group of students, written learner tasks, and unstructured observations. The teacher and I then created materials for four to five lessons using these findings. These design sessions were recorded to take account of the design process and the pedagogical decisions that were made. Subsequently, the teacher employed the intervention in their own CLIL classroom (grade 11), which was videotaped and observed by the researcher. Finally, the process and products of the current cycle were evaluated from the perspective of the learners and the teacher and as indicated in the students' written performances after the intervention. The insights gained then informed the following research cycle, increasingly finetuning the materials and approach developed.

From this process, design principles could be deduced that help teachers create their own content-and-language-integrative materials or adapt the materials designed in this thesis to the needs of their own contexts. Additionally, the data compiled for this study allows a deepening of our knowledge of how content and language learning interlink in the subject history.

### **1.3 Outline**

This thesis is composed of nine themed chapters. Following this brief introduction positioning the empirical study presented in this thesis, chapter 2 is concerned with CLIL as an educational approach and object of scientific interest. This includes a brief exploration of its origin, related labels, as well as its status in Austria, where the study is conducted. Then this chapter zooms in on recent developments in CLIL research relevant for the study at hand, such as outcome-oriented CLIL studies, perspectives of CLIL stakeholders, CLIL materials and didactic design, and the role of diversity in CLIL, concluding with a discussion of how these research strands connect in the light of the present study, i.e., the integration of content and language learning.

Chapter 3 then dives into various conceptualizations of integrating content and language and how these theoretical frameworks have been operationalized for pedagogical use in the subject history. This exploration includes sociocultural theory (SCT), systemic functional linguistics (SFL), the 4Cs framework and its elaboration the pluriliteracies model, and, finally, the CDF construct. By reviewing these concepts and discussing their applicability in the history classroom, the linguistic requirements of historical discourse are outlined from various perspectives in this chapter as well.

Given the transdisciplinary nature of this thesis, chapter 4 is dedicated to the perspective of history education. In this chapter, current central notions and themes in history didactics are reviewed, such as *historical competence*, *historical thinking*, *historical consciousness*, *historical literacy*, and *historical reasoning*. Then, this chapter outlines the competency model underlying Austrian secondary curricula, before moving on to explore the broader context of this study. This includes an overview of the Austrian school system, the school type featured in this study, and the role of history education there.

Introducing the empirical study, chapter 5 presents the methodology used in this study. First, general principles and parameters of DBR are reviewed prior to elaborating on the concrete context of the present study. Then, the research design is outlined, including information on the participants, the organization and process of data collection, the research instruments, methods of analysis, and ethical considerations.

Moving on to the results of the study, chapter 6 first briefly outlines central insights of the pilot study, and chapter 7 presents the findings of the main research cycles. This chapter is structured into four typical phases of a DBR research cycle: needs analysis, design of the intervention (where also the didactic designs created are introduced), classroom implementation, and evaluation of the interventions.

Chapter 8 then discusses these findings against the backdrop of the literature presented in chapter 2, 3, and 4. Here, answers to the research questions that guided this thesis are presented. Moreover, it revisits the central theoretical concept of this dissertation, i.e., the CDF construct, discussing the theoretical implications of this study. Moreover, chapter 8 also outlines methodological insights gained in the course of this PhD project.

Finally, chapter 9 offers a conclusion, summarizing the work that has been done and highlighting key findings and implications of this study. Moreover, this chapter discusses the significance of this study but also its limitations before concluding with recommendations for future research.

## **1.4 Key terms**

Before moving into the main body of this thesis, the use of some terms needs to be clarified. First of all, as chapter 2 will discuss in more detail, there are several labels for situations when learners use another language than their first in lessons scheduled as content subjects, one of which is *CLIL*. When reviewing literature, I will use the same labels as the authors whose work I am quoting. When talking to or interviewing stakeholders, I will use the labels that they are familiar with, i.e., the label used for their programme. However, given the thrust of this study, for my own conceptualization, I will use the label CLIL with an underlying understanding that this is an educational approach where an L2 is not just the medium of instruction but a central element in constructing subject-specific knowledge and skills; the nature and parameters of this integration will be the matter in question of this thesis.

This brings us to another term that needs clarification. I will use the term *L2/ second language* for any language learned after the first language(s). As such, the term *L2* also applies in contexts where an additional language is learned as a *foreign language* (FL). Considering that especially English permeates different aspects of life in countries like Austria despite English not being an official second language (Smit & Schwarz, 2019), the difference between second language and foreign language has been found to be increasingly blurry (e.g., Pecorari, 2018) in any case. I also agree with Schwarz (2020) that it makes sense to use the label *L2* rather than *Lx*, even though *Lx* has been proposed as a more inclusive and value-free label (Dewaele, 2017), simply because *Lx* seems not as commonly used. For example, leading SLA researchers, as for instance the Douglas Fir Group, propose and use the *L2* label (Duff & Byrnes, 2019; The Douglas Fir Group, 2016).

Following this, it also needs to be clarified whether I distinguish between the terms *acquisition*, traditionally implying a more implicit process, *learning*, often assumed to entail deliberate pedagogical attention to features of a language, and resultant *development* (e.g., by Krashen, 1987, see also Lopez Ornat, 2012). However, such a differentiation is difficult to validate empirically in spite of advancements in neuroimaging and neurolinguistics (Roberts et al., 2018). Thus, I will use these three terms interchangeably to describe consciously induced and implicit gains for the domains in focus of this study, i.e., academic and subject-specific CDF use and history competences. The term *competence*, in turn, will be in line with Weinert's definition, which equates competence with "combinations of those cognitive, motivational, moral, and social skills available to (or potentially learnable by) a person or a social group [...] through appropriate understanding and actions of a range of demands, tasks, problems, and goals" (Weinert, 2001, p. 2433). Yet, as pointed out by Pandel (2017) or Klieme and Leutner (2006), such competences are understood to be domain specific. Further elaborations on this topic for the subject history will be provided in chapter 4.

## 2. Content and Language Integrated Learning (CLIL)

### 2.1 Origins and labels

*“One language sets you in a corridor for life. Two languages open every door along the way.”*

Frank Smith (1992, cited in Dzik, 2020, p. 165)

As illustrated by this quote by psycholinguist Frank Smith, being bilingual opens up opportunities, especially in a globalised, all-connected world. In the 1990s, this need for plurilingualism was recognised by the European Union, demanding more intensive and more effective language pedagogy (European Commission, 1995; Mehisto et al., 2009; Perez Vidal, 2009). More specifically, in a white paper on education, the European Commission recommended that “pupils should study certain subjects in the first foreign language learned” so that “everyone should be proficient in two Community foreign languages” (p. 44). In other words, European policymakers strived for fostering multilingualism and increasing students’ linguistic repertoire. Additive bilingual education, i.e., situations where the learning of a second language is seen as an add-on without impeding L1 development, was regarded as a viable way to reach this goal (Dalton-Puffer, 2017). This type of bilingual education (BE) became known as *Content and Language Integrated Learning*, abbreviated as CLIL, and has since spread all over Europe (Dalton-Puffer et al., 2010; Eurydice, 2006, 2017; Pérez Cañado, 2012). In fact, all but four European countries offered some sort of CLIL provision on primary, secondary, and/or tertiary level in 2016/ 2017 (Eurydice, 2017). According to Coyle et al. (2010), the driving forces behind growing interest in CLIL are both reactive, i.e., counteracting existing linguistic and didactic shortcomings, and proactive, as parents wished for providing their children with better socio-economic prospects, educators hoped to improve learning settings, and policymakers wanted “to lay the foundation for greater inclusion and economic strength” (p. 8).

While being a widespread phenomenon, local CLIL practice is dependent on its educational context, rendering CLIL realities multi-faceted and complex. Despite having secured its place within the European educational landscape, top-down policies concerning implementation and quality control of CLIL programmes are rather the exception than the rule, as CLIL programmes are often the result of grassroots movements initiated by committed (head) teachers and parents (Dalton-Puffer, 2017). To complicate matters, there are numerous international, national, or regional labels used for similar conceptualisations which, at their core, share that content and language should be learnt simultaneously. In this vein, CLIL functions as an umbrella term which describes “a dual-focused educational approach in which an additional language is used for the learning and teaching of both content and language” (Marsh & Langé, 2000, p. 2).

While the term CLIL originated in the European context, programmes as described above can be found all over the world (Morton, 2020; San Isidro, 2019), and especially in Asia or South America CLIL programmes have become increasingly popular (Dalton-Puffer, 2017; Pérez Cañado, 2020).

In the North American context, other forms of additive BE have been implemented much earlier, starting with French *immersion* programmes in Canada in the 1960s and spreading to the United States in somewhat different shapes, termed as *content-based instruction (CBI)* or *dual-language/ two-way immersion (TWI)* programmes (Abello-Contesse, 2013; Genesee, 2013; Kim et al., 2015). CLIL, as a descendant of North American bilingual programmes (Pérez Cañado, 2012), shares some of their elements, as they all combine language and content teaching/ learning in one way or another; yet the different programmes vary in their realization, contextual characteristics, theoretical underpinnings, and/ or strategic aims. For instance, while immersion or TWI programmes often combine majority and minority languages with the help of native speaker teachers, CLIL usually targets a (prestigious) FL, which is English in most cases, and tends to be taught by non-native speakers (Dalton-Puffer, 2008; Lasagabaster & Sierra, 2010; Olsson, 2015). Content-based instruction (CBI), typically, describes programmes where curricular content is taught in the official language of a country to learners who, by and large, have not yet fully mastered this majority language (Dalton-Puffer, 2008). Another similar term often used in practice and research is *English-medium instruction (EMI)*. In EMI programmes, the use of English is understood as only having a vehicular function, implying that teachers do not usually deal with the connection between content and language learning, whereas the label CLIL would suggest a fusion of content and language learning goals (Ament & Pérez-Vidal, 2015; Breeze & Dafouz, 2017). These different objectives make EMI more prone to the tertiary sector as means to enable internationalization and attract students (Richter, 2019; Smit & Schwarz, 2019).

## 2.2 CLIL implementation in Austria

Against the backdrop of European policies promoting BE, Austrian CLIL programmes came into being mostly as grassroots movement, often at the initiative of active head teachers and educators, reflecting local needs and possibilities (Hüttner et al., 2013). For quite a long time, there were no official guidelines or regulations regarding the extent or quality of CLIL implementation, resulting in a diverse CLIL landscape mostly managed at the school level (Dalton-Puffer et al., 2018; Hüttner et al., 2013). These local initiatives ranged from small-scale modules to whole bilingual streams (Hüttner et al., 2013).

In addition to local projects, regional educational boards have offered some support and/ or guidelines. In Vienna, for example, the Dual Language Programme (DLP) provides selected secondary schools with English native speaker teachers while leaving the exact parameters of CLIL implementation to the different schools (Vienna Board of Education, 2016). Another programme is called Vienna Bilingual Schooling (VBS), which is a school pilot project where 50% of all lessons should be taught in English, both at primary and secondary level (Vienna Board of Education, 2020). On the part of the ministry of education, who generally appeared to be in favour of bilingual education (Dalton-Puffer et al., 2011), no guidelines were issued for a long time save for a general recommendation that FLs may be used in content subjects (see Austrian Federal Ministry for Education, 2004). Moreover, at primary level, the curriculum calls for integrating English into content subjects, especially in grade 1 and 2, yet without using the label “CLIL”

(Austrian Federal Ministry for Education, 2005). Turning to upper secondary education, in 2011, it was decided that in secondary technical colleges, i.e., vocational upper secondary schools with a focus on technology, 72 hours of CLIL provision across all subjects are obligatory for year 3 to 5<sup>1</sup> (grade 11-13) and recommended in year 1 and 2 (grade 9 and 10, Austrian Federal Ministry for Education, 2015a) to increase employability of the graduates of these schools (Smit & Finker, 2018). This decree was extended to other 5-year-VET schools with different vocational focal points, such as secondary colleges of business administration (Austrian Federal Ministry for Education, 2014) or secondary colleges for agriculture and forestry<sup>2</sup> (Austrian Federal Ministry for Education, 2017b). In schools without mandatory CLIL lessons, new curricula recommend the use of CLIL methodology “to account for social and global developments” (Austrian Federal Ministry for Education, 2015c, p. 12, translated by the author) but do not specify a minimum number of CLIL lessons (see, e.g., Austrian Federal Ministry for Education, 2015b, 2015c, 2018a).

Apart from these general recommendations and, in some cases, regulations on the amount of CLIL lessons or guidelines for teachers (see, for instance, Austrian Federal Ministry for Education, 2017b), no further top-down specifications in terms of concrete implementations have been issued. Nonetheless, a number of common features of Austrian CLIL practice can be observed, most of which are quite typical for European CLIL practice in general. For example, Austrian CLIL education is usually content-driven, as is the case in most European countries (Dalton-Puffer et al., 2010). To be more precise, CLIL classes usually adhere to the curriculum of the content subject, are taught by content specialists, and are assessed via subject-specific criteria (Nikula et al., 2016). Like in most European settings, Austrian CLIL lessons are timetabled as subject classes and thus do not replace FL teaching. Such a conceptualization of CLIL is often described as *hard CLIL* or *Type A CLIL* as opposed to language-driven *soft* or *Type B CLIL*, which is integrated in the FL programme (Dalton-Puffer, 2017).

Another feature Austrian CLIL practice shares with other contexts is the predominance of English as foreign language (EFL) (Eurydice, 2017; Smit & Schwarz, 2019). There are some few schools that offer bilingual programmes with Romance languages, Chinese, Polish, Arabic, or regional minority languages, such as Croatian, Hungarian, or Czech (Eurydice, 2017), but English is, by far, the most common CLIL language in Austria (Smit & Schwarz, 2019).

Turning to Austrian secondary CLIL teachers, again there are no official regulations concerning the qualification and selection of CLIL teachers (Austrian Federal Ministry for Education, 2017b). Unlike in many other educational systems, Austrian secondary teachers have to qualify for two subjects (with a few exceptions such as religious education teachers). Unsurprisingly, teachers offering a combination of English and another content subject are often key figures in local CLIL programmes. As for pre- and in-service teacher training, CLIL has become a more central topic, but these courses are not obligatory as a rule. In-service teacher training is available in different forms and scales, ranging from short training sessions focused on one aspect of CLIL to

---

<sup>1</sup> In year 5, only 40 lessons are obligatory.

<sup>2</sup> Here, only 36 CLIL lessons a year are mandatory.

comprehensive courses consisting of various modules over a number of semesters (see, for instance, KPH Vienna/ Krems, 2020; Teacher Training College Upper Austria, 2020; Teacher Training College Vienna, 2020). In the end, the decision who teaches CLIL is made on the local level. Some head teachers expect their CLIL teachers to complete a course or attend in-service training; others insist on previous experience abroad or a recent language certificate in case they are not qualified English teachers. In some cases, and as sanctioned by the ministry, “enjoying languages” suffices (Austrian Federal Ministry for Education, 2017b, p. 13, translated by the author).

## **2.3 Recent developments in CLIL research**

As already indicated above, CLIL is a widely implemented and “multi-faceted phenomenon” in practice (Nikula, Dalton-Puffer, & Llinares, 2013, p. 74), which inspired a very active and lively research scene (Hüttner & Smit, 2014). This research field is explored below, starting with a brief overview of North American research on BE (2.3.1). Then, main strands of CLIL research relevant for this dissertation are concisely summarized, which includes outcome-oriented studies (2.3.2), participants’ perspectives (2.3.3), materials and didactic practices (2.3.4), and diversity in CLIL (2.3.5). Considering the plethora of CLIL research and the relatively fast pace of change within this field, this part focuses on research conducted in the last seven years, but, if relevant, earlier studies are considered too. Obviously, there are also other substantial sub-fields within CLIL research, such as studies into classroom discourse and translanguaging, assessment, or policies with regards to CLIL, but these are not as pertinent to the dissertation at hand and therefore are not featured below. Finally, this subchapter concludes with an attempt at linking the four strands presented in this literature review (2.3.6).

### **2.3.1 Looking back: research on Canadian and American bilingual education**

Given their conceptual similarities, the myriad of research on other forms of additive BE, such as immersion or CBI, has played a crucial role for the comparatively young CLIL research tradition, adding other perspectives, experiences, and previous insights to the field (Dalton-Puffer, Llinares, Lorenzo, & Nikula, 2014). Looking at literature reviews by Pérez Cañado (2012), Tedick and Wesely (2015) and Genesee (2006), central threads in North American CBI/ immersion research concern the impact of BE on learners’ L2 and L1. Most studies conclude that immersion learners usually reach near-native listening and reading skills, but productive skills do not develop to the same extent, while the children’s L1 development is not negatively affected (Genesee, 2006; Pérez Cañado, 2012; Tedick & Wesely, 2015). On the affective level, most studies show that immersion learners view the target language and its speakers positively (Pérez Cañado, 2012) and demonstrate higher levels of cross-cultural competence and satisfaction with their own education (Tedick & Wesely, 2015). Finally, no negative effect could be observed in terms of content learning (Genesee, 2006; Pérez Cañado, 2012; mathematics: Tedick & Wesely, 2015) and general cognitive development (Pérez Cañado, 2012).

While research on Canadian immersion and American BE offers some groundwork and valuable perspectives, Pérez-Cañado (2012) stresses that one needs to keep in mind that North American research outcomes “cannot be simply transferred or transposed to the European scenario” (p. 318) due to their contextual differences. For this reason, research on CBI or immersion will not be further explored here more generally but will be mentioned in other parts of the dissertation where appropriate. As such, these might link to various aspects North American research on BE has thoroughly examined, including pedagogical practices, teacher development, or material design.

### 2.3.2 Outcome-oriented CLIL studies

Similar to immersion/ CBI research, CLIL research has put emphasis on outcome-oriented studies, comparing performances of mainstream and bilingual students. Although CLIL practice tends to be defined by content-related concerns, the research community has, without doubt, focused on linguistic outcomes, whereas content-related aspects have been studied far less and have only recently started to receive more attention (Dalton-Puffer, 2018; Fernández-Sanjurjo et al., 2017; Morton, 2020; San Isidro, 2019). One reason for this language-bias within CLIL research is that most CLIL researchers are linguists and are thus, naturally, more interested in linguistic aspects (Dafouz et al., 2014). Content specialists, on the other hand, often appear sceptical of the surplus value of CLIL (Piesche et al., 2016; Theis, 2010) and fear that the use of an FL might interfere with depth and breadth of content learning (Dalton-Puffer, 2008) and could also overburden both students and teachers (Heimes, 2011). At the same time, eclipsing content specialists from CLIL research would not only lead to an incomplete representation but also to disadvantages for teacher professionalism (Maset, 2015), teacher education, and in-service teacher support (Cammarata & Cavanagh, 2018; Lo & Jeong, 2018; Morton, 2018).

The numerous studies looking into the impact of CLIL on linguistic performance have often reported mostly positive effects of CLIL on the different aspects of language acquisition (Dalton-Puffer, 2008; Goris et al., 2019; Olsson, 2015; Pérez Cañado, 2012; San Isidro, 2019). However, many outcome-oriented studies contributing to CLIL’s ‘good reputation’ have recently been viewed more critically (Bruton, 2011; Fernández-Sanjurjo et al., 2017; Pérez Cañado, 2016b; San Isidro, 2019). One area of criticism stems from the reality that many CLIL programmes select students with high levels of linguistic aptitude, motivation, and/or socio-economic backgrounds, leading to an elitist notion of CLIL (Broca, 2016; Bruton, 2011, 2013, 2017; Graham et al., 2018; Paran, 2013). Many (early) studies did not take selection biases into account, resulting in unsuitable matching of control groups (Bruton, 2011; Pérez Cañado, 2016b; Piesche et al., 2016).<sup>3</sup> Another central point of criticism concerns the statistical analyses of comparative outcome-oriented CLIL studies. It has been pointed out by several researchers (e.g., Graham et al., 2018; Meyerhöffer & Dreesmann, 2019; Pérez Cañado, 2012, 2016b; Piesche et al., 2016; Roussel et al., 2017) that some previous studies did not include pre-tests (e.g., Dafouz et al., 2014; Ouazizi, 2016;

---

<sup>3</sup> Examples: Admiraal et al. (2006); Lasagabaster (2008); Ruiz de Zarobe (2007)



Villarreal Olaizola & García Mayo, 2009) or relevant statistical measures when analysing outcomes (e.g., Alonso et al., 2008; Ouazizi, 2016) or partly applied insufficient or inadequate statistical procedures (e.g., Admiraal et al., 2006; Rosi, 2018). As a consequence, the effectiveness of CLIL has recently been viewed more sceptically (Pérez Cañado, 2016b; San Isidro, 2019). Moreover, it has been pointed out that there are too few longitudinal studies, raising questions about the long-term impact of CLIL (Pérez Cañado, 2012; San Isidro, 2019). Recently, more emphasis has thus been put onto longitudinal, comparative studies, and/or outcome-oriented studies that aim for adequate matching of control groups and statistical analysis. Compared to initial studies yielding mostly positive results, the outcomes of these studies have been more mixed (e.g., Dallinger et al., 2016; Gierlinger & Wagner, 2016; Roquet & Pérez-Vidal, 2017; Rumlich, 2016, see literature reviews by Goris et al., 2019; Graham et al., 2018).

Looking at various aspects of language learning in a little more detail, most studies on reading skills suggest a positive impact of CLIL (e.g., Admiraal et al., 2006, Artieda et al., 2017; Pérez-Vidal & Roquet, 2015; Prieto-Arranz et al., 2015; Sylvén & Ohlander, 2019a). Yet, it should be noted that in Artieda et al.'s (2017) study, CLIL students significantly outperformed their peers only when matched for age and not when matched for hours of exposure, indicating that a certain threshold of exposure time would be needed to harness a positive effect.<sup>4</sup> Here, a number of intervention studies suggest that the implementation of reading strategies might be key for gaining better reading comprehension skills in CLIL (e.g., Bayram et al., 2019; Hamidavi et al., 2016; Quintana Aguilera et al., 2019; Ruiz de Zarobe & Zenotz, 2018). Zooming in on the type of reading skills, Prieto-Arranz et al.'s (2015) findings showed that the CLIL learners significantly outperformed their EFL peers in the specific reading comprehension tests at all testing times, whereas in the general reading part of the test the differences were not statistically significant, and at one testing time (out of four), the CLIL learners were surpassed by the EFL group. More mixed results were further reported in Goris et al. (2013), while neutral effects were observed in Pladevall-Ballester and Vallbona (2016) and its follow-up study by Pladevall-Ballester (2016), showing that CLIL science students had an advantage over arts and crafts CLIL learners.

Studies on listening skills, overall, are comparatively scarce and do not represent a clear picture either (Pérez Cañado & Lancaster, 2017; Prieto-Arranz et al., 2015). For example, Pérez Cañado and Lancaster (2017) observed a positive effect for oral comprehension, but this effect was smaller than for oral production. Dallinger et al. (2016) also found a positive impact on listening skills of CLIL learners, while their general English skills were unaffected. Prieto-Arranz et al. (2015), on the other hand, reported mixed results, and Pérez-Vidal and Roquet (2015) described a neutral effect of CLIL on listening comprehension, while the control group in Pladevall-Ballester and Vallbona's (2016) study even significantly outperformed the CLIL cohort in terms of listening skills, again showing that CLIL science learners presented better results than their arts and crafts CLIL peers. While the studies mentioned above do not detail the type of listening skills, Nieto

---

<sup>4</sup> The authors suggest 280-300 hours.

Moreno de Diezmas (2018) investigated various subskills of listening. Her results suggest that at secondary level, CLIL learners outperformed their EFL peers in all subskills, whereas CLIL learners at primary level only scored significantly higher in terms of global comprehension and details but not concerning understanding the situation, paralinguistic elements, vocabulary, or space-time relations.

Turning to productive skills, most studies have shown that speaking skills are positively affected by CLIL (e.g., Admiraal et al., 2006; Mewald, 2007; with matched groups: Gallardo del Puerto & Gómez Lacabex, 2017; Pérez Cañado & Lancaster, 2017; San Isidro & Lasagabaster, 2019). Moreover, focusing on oral narratives, several studies have demonstrated advantages of CLIL cohorts over their mainstream peers (e.g., Gallardo del Puerto & Gómez Lacabex, 2017; Hüttner & Rieder-Bünemann, 2010; Martínez Adrián & Gutiérrez Mangado, 2015). Yet, in Gallardo del Puerto and Gómez Lacabex's (2017) study, not all of the differences reported reached statistical significance, and the sample in Martínez Adrián and Gutiérrez Mangado (2015) was very small. The study by Hüttner and Rieder-Bünemann (2010) reported qualitative differences between the groups. In terms of turn-taking and cooperation, several studies observed a positive CLIL effect. For example, P. Moore's (2011) results suggest that CLIL learners collaborate more often and more effectively. Similarly, Pastrana et al.'s (2018) findings indicate that CLIL students cooperate better and, in collaboration, engage with the content to a greater degree. Mesquida and Juan-Garau (2013), too, found that CLIL learners' negotiation strategies were better developed at each collection time, but the authors added that the CLIL group received more exposure to English than their peers. In the same setting, the pronunciation of CLIL students was less accented and more intelligible (Rallo Fabra & Juan-Garau, 2011). In other studies, however, accent (as well as some other measures of fluency) were not affected (Rallo Fabra & Jacob, 2015), unless there was a phonetic intervention helping learners improve their pronunciation (Gómez Lacabex & Gallardo-del-Puerto, 2020).

As for written production, the results of previous studies suggest a mostly positive impact of CLIL provision on the students' writing (e.g., Gené-Gil et al., 2015; Lahuerta, 2015, 2020; Pérez-Vidal & Roquet, 2015; Ruiz de Zarobe, 2010). Dalton-Puffer and Jexenflieger (2010) also observed a positive effect on the students' writing in terms of task fulfilment, grammar, and vocabulary but not for organisation and structure. Having matched groups for exposure time, Artieda et al.'s (2017) findings were more mixed, as CLIL students only outperformed their peers when matched for age in terms of lexical richness as well as linguistic and communicative competence. When matched for exposure time, however, the older, non-CLIL students' writing demonstrated significantly higher levels of accuracy and coordination. Similarly, in a study by Roquet and Pérez-Vidal (2017), where groups roughly had the same amount of exposure time, CLIL students only surpassed their peers in terms of accuracy but in no other domain under investigation. Thus, again it seems that cognitive maturity might be more important than type of instruction. Looking into pragmatic competence, age and length of instruction appear to play a bigger role than whether learners were in CLIL programmes or not. For example, analysing written requests by CLIL

learners and older EFL-only students with extra exposure to English, Nashaat Sobhy (2017) found that the latter performed better than the former. At tertiary level, too, Codina-Espurz and Salazar-Campillo (2019) observed that both high- and low-intensity CLIL learners struggled with producing adequate e-mail requests, yet the high-intensity students presented a greater range of moves.

The effect of CLIL on vocabulary learning has been studied quite extensively, with the majority of studies indicating a positive impact of CLIL on the learners' lexis. In terms of receptive vocabulary, a great number of studies have found that CLIL learners significantly outperform their mainstream peers (e.g., Agustín-Llach & Canga Alonso, 2016; Bayram et al., 2019; Canga Alonso, 2015b; Castellano-Risco, 2018; Castellano-Risco et al., 2020; Castro-García, 2017; Iglesias Diéguez & Martínez-Adrián, 2017; Sylvén & Ohlander, 2019b; Xanthou, 2011). Here, it should be kept in mind that in some of these studies, exposure time has not been factored in. Those studies that (additionally) compared older EFL-only learners with younger CLIL learners who received similar hours of exposure often reported neutral effects (e.g., Agustín-Llach, 2015; Arribas, 2016; Canga Alonso, 2015a; Fernández Fontecha, 2014; Fernández-Fontecha, 2015; Goris et al., 2013<sup>5</sup>; Iglesias Diéguez & Martínez-Adrián, 2017; Verspoor et al., 2015). Interestingly, some of these studies reported that learners' motivation was a predictor for vocabulary size rather than type of instruction (Arribas, 2016; Fernández Fontecha, 2014), while the findings by Castellano-Risco et al. (2020) suggest that type of instruction plays a bigger role than exact hours of exposure. Gierlinger and Wagner (2016) also observed a zero-effect overall, but they add that growth could only be found in the K1 dimension, i.e., the 1000 most frequent words in the English language. Therefore, they call for a more deliberate approach to vocabulary learning in CLIL which also considers academic and subject-specific words. Moving on to productive vocabulary, positive but, by and large, rather small effects have been reported in Bayram et al. (2019), Crossman (2018), Jiménez Catalán and Agustín-Llach (2018), or Navarro Gil (2019). Very young CLIL learners, however, do not seem to be at an advantage (Agustín-Llach, 2016; Agustín-Llach & Jiménez-Catalán, 2018). Looking at academic vocabulary, Olsson's (2015) findings indicate a neutral effect if individual differences are factored in. In fact, Olsson and Sylvén (2015) could show that a high level of extramural exposure to English correlates with a slower growth in terms of academic vocabulary, and CLIL learners often tend to be in more contact with English outside the classroom.

Finally, (lexico-)grammar has received less attention, but the few studies conducted point towards a positive effect. For example, Juan-Garau et al. (2015) reported that CLIL learners presented a steeper learning curve than their EFL-only peers. When matched for age, Artieda et al. (2017) also observed a CLIL advantage, but when matched for hours, the groups were rather similar in their knowledge about grammar. Taking a longitudinal perspective, Lorenzo et al.'s (2019) findings show that over the course of three years, CLIL learners improved statistically significantly in terms of complexity and syntactic pattern density but not in other domains such as cohesion or

---

<sup>5</sup> Neutral effects were reported for the German and Dutch data set; in the Italian data, a positive impact was observed.

use of connectives.<sup>6</sup> The CLIL learners in Bulon's (2020) study, however, presented a greater phraseological variety and, partly, accuracy. Similarly, the results by Möller (2017) indicate a greater variety and accuracy of passive constructions in writings by a group that opted for CLIL compared to a group that deliberately avoided CLIL and a group where CLIL was not available (but where the context was still considered selective). Here, it is interesting to note that the difference between the voluntary CLIL group to the selective non-CLIL group was smaller than to the group actively not choosing CLIL.

Moving on to studies focusing on the impact of CLIL on content learning, most studies seem to suggest a zero effect (Dalton-Puffer, 2008; Meyer et al., 2015; Pérez Cañado, 2012; San Isidro, 2019). For example, Badertscher and Bieri (2009) investigated secondary bilingual geography, history, and biology teaching in Switzerland and conclude that content learning does not suffer, but bilingual learners might need both languages to be able to demonstrate what they have learned. Similarly, Gablasova (2014), who looked into accuracy, fluency, and lexical and academic appropriateness of historical definitions, found that bilingual learners struggled with declarative content knowledge transfer from L2 back to L1. Admiraal et al. (2006), apart from examining CLIL effects on the learners' linguistic skills, compared final history and geography grades of bilingual and mainstream students and found no significant difference. Similarly, a large-scale study in Germany by Dallinger et al. (2016) on language learning and factual historical knowledge reported no differences in the history scores of the CLIL and the control group, although the bilingual learners had one more history lesson a week, indicating that the bilingual students progressed more slowly. The authors, however, add that a broader understanding, i.e., one that encompasses historical skills, might have resulted in a more positive outcome, as historical learning is assumed to benefit from the use of foreign languages (e.g., increased ability to take over other perspectives, see subchapter 4.2 for more information).

In science education, most studies, too, observed zero effects. For instance, in Xanthou's (2011) study on the effect of CLIL in primary education in Cyprus, the experimental and control groups' demonstration of science knowledge developed quite similarly. A quasi-experimental study by Meyerhöffer and Dreesmann (2019) also indicates a neutral effect on biology learning in secondary education, but the experimental group perceived the intervention as positive and wished for more CLIL in the future. Fernández-Sanjurjo et al. (2017) compared CLIL and non-CLIL students' science knowledge at the end of Spanish primary education (grade 6) and found that the students taught in the L1 performed better than their bilingually taught peers. Fernández-Sanjurjo et al. (2017) argue that better teacher training and support might be key to ensure that CLIL learners do not fall behind. Another interesting insight of their study is that socio-economic status indeed was a determining factor for the students' performance. Looking at age as a variable, Hughes and Madrid (2020) found that the scores of primary CLIL and non-CLIL groups did not

---

<sup>6</sup> In the copy-editing phase of this PhD project, Granados et al. (2021) published a follow-up study working with the same corpus, comparing the learners' development in their L2 (English) and their L1 (Spanish). The results of this study point towards a parallel development in a number of dimensions, including nominalisation, length measures, subordination, and lexical development.

differ, while secondary CLIL students outperformed their peers regarding knowledge about science.

All these studies conceptualize content learning as declarative, factual knowledge. This, however, does not fully represent the current educational landscape, as competency-based curricula have replaced, or at least expanded, traditional knowledge-based curricula in many countries (Gautschi, 2015; Priestley & Biesta, 2017). Piesche et al.'s (2016) experimental study in the CLIL subject physics seems to assume a somewhat semi-competency-based approach, as their test, despite targeting reproductive knowledge for the most part, includes a competency-based free-response item too, and their intervention itself appears to be competency-based and constructivist, where learning is seen as a learner-centred, active process that includes hands-on experiments. In this study, the bilingual students produced worse results than the control group both in the immediate and delayed post-test. The authors argue that the students had no previous experience with bilingual teaching and might thus have been overwhelmed with the situation. Moreover, Piesche et al. (2016) hypothesize that the students' language level might have been too low, resulting in an overload of working memory capacity (see cognitive load theory, Sweller, 2011). This somewhat ties in with a study by Fung and Yip (2014), who investigated the impact of EMI instruction in upper secondary science education with regard to different achievement levels. Testing students' knowledge and problem-solving skills after an entire academic year, Fung and Yip (2014) found that low-ability students benefitted more from L1 instruction, whereas high achievers attained better results in the EMI setting.

In social sciences, San Isidro and Lasagabaster (2019), who also seemed to follow a competency-based approach, observed a zero effect. In a large-scale study by Pérez Cañado (2018b), which tested both declarative and procedural knowledge, there also was a neutral effect on primary level, but CLIL students on secondary level surpassed their peers in terms of science learning. This seems to correspond to the results by Hughes and Madrid (2020) mentioned previously.

Turning to mathematics, Jäppinen (2005) compared the thinking and learning processes of experimental CLIL and non-CLIL students in three different age groups. Overall, the author concludes that a CLIL environment supports the cognitive development, especially for the age group 10-12. The oldest group (13-15), however, showed almost no difference, while younger learners (7-9) seemed to struggle with abstract concepts. A study by Surmont et al. (2016) seems to confirm the results for secondary learners (aged 12, approximately) of mathematics who just started with CLIL instruction. These learners outperformed their traditionally taught peers after three months of CLIL instruction and even more so after 10 months. The authors explain these results with a more effective cognitive stimulation, metalinguistic awareness, and deeper processing (see also Surmont et al., 2014).

### **2.3.3 Participant perspectives**

Another early avenue of CLIL research concerns the attitudes, beliefs, and perspectives of CLIL learners and teachers.

### 2.3.3.1 Affective factors

Among affective factors, learner motivation has been a central topic in CLIL research, with most studies reporting a positive relationship between CLIL and level of motivation (Banegas & Pinner, 2021; Coyle, 2013; San Isidro, 2019; Somers & Llinares, 2018). However, Coyle (2013) argues that these studies are difficult to compare, as there are so many variables at play (see also Doiz et al., 2014). Concerning attitudes towards learning an FL, the results are, by and large, consistent, indicating that most CLIL learners are more motivated to learn and use the target language (e.g., Lasagabaster & Sierra, 2009; Doiz et al., 2014; Lasagabaster, 2011; Mearns, Graaff, & Coyle, 2017; Merisuo-Strom, 2007; Verspoor et al., 2015). These positive attitudes seem to prevail in the long run (e.g., Pladevall-Ballester, 2019) and even after learners have finished school (Roiha, 2019). Yet, it needs to be kept in mind that motivation can vary throughout a student's learner biography and that positive attitudes and high levels of motivation concerning language learning might have been present even before learners start BE, leading to opting for CLIL in the first place (Dallinger et al., 2018; Mearns et al., 2017; Rumlich, 2016). Lasagabaster and Doiz's (2015) results, for example, underscore that in their longitudinal study on affective factors in CLIL, age as well as whether learners were selected into the CLIL programme or not were determining factors for the learners' motivational development. While the non-CLIL learners could sustain their motivation, CLIL learners started with high levels of motivation, which later decreased somewhat. Interestingly, the younger group, which was also the group that was selected to participate in CLIL, was affected by this trend to a greater extent than the older CLIL students who did not experience selection. In the Austrian context of mandatory CLIL education in vocational upper secondary education, Döring (2020) reports that almost half of the participants of her case study said that they were not more motivated to use English in CLIL lessons than in their traditional EFL classes, in contrast to 36% who felt more motivated. Interested in avoiding selection bias, Ohlberger and Wegner (2017) compared the motivational development of regular bilingual classes, functioning as control group, to those that had never had CLIL instruction before, i.e., the treatment group. Their results showed that neither group was more motivated towards the use of English after the intervention than before, countering the claim that CLIL instruction per se would lead to more motivated students. Yet, the authors argue that the novelty of the approach might have overburdened some of the learners of the treatment group, preventing an increase in motivation. Considering gender as a variable, results by Heras and Lasagabaster (2015) indicate that CLIL instruction redressed gender-related differences in motivation, but overall they could not identify significant differences in motivation between CLIL and mainstream learners.

Looking at other affective factors, CLIL has been found to be linked to lower levels of anxiety (Simons et al., 2019; Thompson & Sylvén, 2015). In fact, the learners in the study by Milla and García Mayo (2021) did not seem concerned with feelings of anxiousness and even wished for more oral corrective feedback in the classroom. Looking at EFL confidence and motivation, Banegas and Pinner (2021) observed that the synergy between learning content and language has boosted these affective factors in student-teachers who completed a sociolinguistics CLIL module.

At secondary level, Goris et al. (2017) found a positive trend for both CLIL and non-CLIL learners with regards to EFL confidence. Moreover, a longitudinal case study by Roiha and Mäntylä (2021) points towards a positive long-term influence of CLIL on learners' self-concept. As for affectivity and intellectual helplessness, Otwinowska and Foryś (2017) found that CLIL students (grade 4 and 5) experienced more negativity than positivity. Moreover, one in five students scored very high on the Intellectual Helplessness Scale and almost 70% reported some of the symptoms of intellectual helplessness, indicating cognitive overload and feelings of exhaustion. Interestingly, only the students' CLIL subject grades (mathematics and science) were predictors for intellectual helplessness while English grades were not (Otwinowska & Foryś, 2017). This suggests that the reason for feelings of exhaustion and frustrations might lie in the demands of the subject, or, as Otwinowska and Foryś (2017) argue, in the use of cognitive, academic language (which is not present at this stage of English instruction). The authors therefore call for better language support and scaffolding as well as for EFL teaching that goes beyond interpersonal communication skills.

In short, while on first glance, there seems to be a positive link between affective factors and CLIL, a closer look suggests that this relationship has not been clearly established. What is more, those studies that report neutral or even negative effects tend to mention feelings of frustration and overextension, which could be addressed with better scaffolded CLIL materials and teacher training that sensitizes teachers towards these issues. Therefore, Somers and Llinares (2018) argue that such studies should not be understood as evidence against CLIL but rather as an incentive to improve CLIL provision and teacher support. Another issue reported by Somers and Llinares (2018) concerns how motivation in CLIL is conceptualized, namely as solely focused on language learning without considering how this relates to content learning. Instead, they propose to look at *CLIL motivation*, following a language-and-content-integrative approach that should be inherent to this educational model. Comparing motivation of CLIL students in Spain in low- and high-intensity tracks (reflecting their levels of ability), Somers and Llinares (2018) found that the high-intensity group, which mostly received CLIL education in 'more academic' subjects such as maths or science, was more motivated both intrinsically and instrumentally. Somers and Llinares (2018) explain that this could create a Matthew effect, meaning that highly proficient and motivated students experience success and thus become more proficient and motivated, whereas the low-intensity students, who get fewer opportunities to immerse themselves, do not evolve as much. Another interesting result of their study is that, although the level of anxiety was rather low in general, the students of both groups reported to be less anxious in non-CLIL lessons. In light of these results, Somers and Llinares (2018) recommend providing better language support and scaffolding to reduce anxiety and to create an environment that fosters feelings of success and achievement, irrespective of the learners' level of ability.

### 2.3.3.2 Beliefs and perceptions

Moving on to beliefs and perceptions of participants<sup>7</sup>, the perspective of teachers seems to have received more attention than the students' or their parents' (San Isidro, 2019; Somers & Llinares, 2018). Studies focusing on teachers often describe CLIL practitioners as dedicated, involved, and enthusiastic (O Ceallaigh et al., 2017; Pérez Cañado, 2012) and report that teachers view CLIL teaching as challenging but rewarding and motivating (O Ceallaigh et al., 2017; Pavón Vázquez & Méndez García, María del Carmen, 2017; Pérez Cañado, 2012; Skinnari, 2020). Positive sentiments were found to be connected to a high degree of creative freedom and agency for teachers (Gruber et al., 2020; Pappa et al., 2017a; Skinnari, 2020; Talbot et al., 2021), constructive collaboration amongst colleagues (Pappa et al., 2017a; Pavón Vázquez & Méndez García, María del Carmen, 2017), or the teachers' perception that learners would be more motivated and active in CLIL (Gruber et al., 2020; San Isidro, 2019; Somers & Llinares, 2018). Challenges perceived by practitioners involve

- added workload (Massler, 2012) without recognition (Gruber et al., 2020; Pappa et al., 2017b; Talbot et al., 2021) or financial compensation (Skinnari, 2020),
- time constraints within teaching, i.e., not having enough time to cover all topics (Lo & Jeong, 2018; Massler, 2012),
- unclear guidelines and policies (Skinnari, 2020; Talbot et al., 2021) especially concerning assessment (Morton, 2020),
- a dearth of adequate CLIL materials (Hahn, 2019; Massler, 2012; Meyer et al., 2015; Morton, 2013),
- insufficient pre- and in-service teacher training (Hahn, 2019; Massler, 2012; Pérez Cañado, 2016a),
- and issues with the target language and/or language teaching skills (Gruber et al., 2020; Lo & Jeong, 2018; Massler, 2012; Pérez Cañado, 2016a; Skinnari, 2020).

Focusing on teacher roles and identity, studies have found that in hard CLIL settings, the teachers usually act and see themselves as content teachers, prioritizing content teaching aims (Dalton-Puffer, 2007; Kong et al., 2011; Lo & Jeong, 2018; Milla & García Mayo, 2021; Morton, 2019; Skinnari & Bovellan, 2016; Tan, 2011). In the context of Austria, where many CLIL teachers are qualified in the language and the subject, teachers also prioritize the role of subject teacher, and as Dalton-Puffer (2007) reports, even “confessed to feeling ‘guilty’ about having acted ‘too much like a language teacher’” (p. 5). Similarly, the two focus teachers in a study by Gierlinger (2021), emphasized that they were, in fact, not language teachers but subject specialists who simply were enthusiastic users of English. Following these teachers for over a year, Gierlinger (2021) described the L2 confidence of these two teachers as complex, unstable, and domain-specific. For

---

<sup>7</sup> While many researchers use *beliefs* and *perceptions* synonymously, others understand *beliefs* as pervasive and encompassing cognitions of participants, whereas *perceptions* are assumed to be more limited and to refer to specific experiences (Wesely, 2012).



these reasons, it is not surprising that Austrian CLIL teachers tend to view language learning in CLIL settings as incidental learning, with the exception of vocabulary learning, which tends to be taught explicitly in CLIL (Dalton-Puffer, 2007; Gierlinger, 2021; Hüttner et al., 2013). In general, it appears that BE teachers often have little awareness of how content and language are related beyond the lexical level (Cammarata & Cavanagh, 2018; Lo & Jeong, 2018; O Ceallaigh et al., 2017; Skinnari & Bovellan, 2016). Instead, CLIL teachers often seem to hold the view that CLIL is beneficial for language learning because the students would 'naturally' absorb the language without needing much further attention (Dalton-Puffer, 2007; Hüttner et al., 2013; Skinnari & Bovellan, 2016).

Other studies have reported that CLIL teachers do contemplate about how to balance or integrate content and language learning in CLIL. In the Dutch context, van Kampen et al. (2017) found that practitioners were mostly aware of the role of subject-specific discourse in CLIL, but their conception thereof seemed to be limited to subject-specific terminology too. Likewise, in a case study by Milla and García Mayo (2021), the CLIL teacher recognized the role of language for subject learning and also saw the benefits of oral corrective feedback, yet this did not carry into classroom practice, where this focus teacher hardly corrected, save for lexical errors. Some studies dealing with teachers' beliefs concerning the integration and balancing of content and language learning reported feelings of uncertainty and insecurity towards the professional identity and integrity of bilingual teachers, partly owing to the fact that most of them are either trained content or language teachers (Bonnet & Breidbach, 2017; Moate, 2011; Talbot et al., 2021). Comparing primary, secondary, and tertiary CLIL teachers, Talbot et al. (2021) report that worries about balancing content and language were mainly found among primary teachers, up to a point that would discourage them from teaching CLIL, whereas upper secondary and tertiary teachers felt that language was not their concern and were thus less stressed about this issue and CLIL in general. The teachers surveyed in Hunt (2011) considered this question to be an opportunity to re-evaluate and re-shape one's practice, ultimately improving the learning experience of their pupils.

Studies focusing on students' beliefs and perceptions have mostly shown that students tend to believe in the usefulness of CLIL (Barrios & Milla Lara, 2020; Dalton-Puffer et al., 2021; Massler, 2012) and often have high expectations concerning their future use of English (Broca, 2016; Calderón-Jurado & Garcia, 2018; Oxbrow, 2018). In subjects where a future benefit is not as obvious, e.g., in history, learners often found CLIL less purposeful (Somers & Llinares, 2018). Contrary to fears of parents and teachers, CLIL learners tend to report that they can understand the content taught in CLIL lessons (Massler, 2012; Pladevall-Ballester, 2015) and do not complain about missing depth of content (Dalton-Puffer et al., 2021). However, low achievers stated that they were not always able to follow (Calderón-Jurado & Garcia, 2018; Massler, 2012; Pladevall-Ballester, 2015). In other studies, CLIL learners mentioned that the use of English made things more difficult (Pérez Cañado, 2012), which entails student selectivity (Broca, 2016). It has also been noted in mandatory CLIL contexts that participants do not believe that all subjects are

equally suitable, as in some content areas, the use of an FL overcomplicates learning, especially if the teachers lack L2 proficiency (Dalton-Puffer et al., 2021). Furthermore, it was found that these learners' satisfaction appeared to be contingent on the level of support provided (Dalton-Puffer et al., 2021).

Looking at the distribution of roles and pedagogical practices as perceived by learners, Hüttner et al. (2013) report a "more relaxed" (p. 276), collaborative atmosphere and a more flexible allocation of roles. Their interviews with students reveal that learners feel more eye-to-eye with CLIL teachers, as everybody is essentially an English learner, including the teachers, resulting in collaborative negotiations of meaning and little attention to form. Smit and Finker (2018) in their study on CLIL in secondary technical colleges, where CLIL is now mandatory, observed something similar, which students described as student-friendly, entailing slower pace and more explanation (see also Dalton-Puffer et al., 2021, who analysed a related data set). In the same context, Döring's (2020) results suggest that learners appreciate learning technical terminology in English, as this is something they can put to use in their future careers (see also Dalton-Puffer et al., 2021). To a lesser extent, these learners also valued the possibility to use English in a safe environment (Döring, 2020). In other studies in the Austrian context, such sentiments were reported too, describing CLIL as a highly learner-centred approach that also engages them cognitively (Bauer-Marschallinger et al., 2021; Dalton-Puffer et al., 2021). At the same time, Austrian students argued that CLIL instruction could benefit from better teacher qualification in the L2, from being granted the right to choose CLIL rather than being forced (Dalton-Puffer et al., 2021; Döring, 2020), and from more frequent use of authentic materials (Döring, 2020). The respondents of Oxbrow's (2018) survey of CLIL learners on the Canary Islands were mostly satisfied with the materials used in CLIL, highlighting their authenticity, and appreciated the "task-based and lexically-focused methodology exploiting projects and collaborative learning" (p. 147). Furthermore, these students felt that the language skills of their instructors were adequate. A focus on vocabulary was also found in a study by Barrios and Milla Lara (2020), who reported that students felt that CLIL was most effective for vocabulary learning, reflecting the teachers' frequently reported practice of explicit vocabulary teaching.

### **2.3.3.3 Participants' voices in didactic design**

While it is interesting and crucial to know how CLIL teachers and learners feel and perceive CLIL instruction in general, taking the participants' voices into account when creating didactic materials and tools, e.g., in the context of an intervention-, a DBR-, or an action research study, is assumed to be a key element for the success of the design (Dijkstra et al., 2017; Filice, 2021; Lo & Jeong, 2018; McKenney & Reeves, 2012). In contrast to a descriptive treatment of attitudes, perceptions, or beliefs, the concept of *voice* entails agency, i.e., giving the participants active roles in educational research and reform (see Cook-Sather, 2006, 2020; Groundwater-Smith & Mockler, 2016; Mitra, 2018; Skinnari, 2020).

In DBR (or its relatives, such as collaborative action research), the teacher especially plays a decisive role for the design and implementation of didactic interventions, providing expertise and

ensuring ecological validity and viability (van den Akker & Nieveen, 2016), such as in Banegas (2013), Lo and Jeong (2018), or Tedick and Young (2018). While the legitimacy of considering the teacher's voice seems quite common-sense given their qualification and role, the voices of CLIL students have not really been considered much despite them being the target audience (Filice, 2021). In most cases, students' views are only considered after an intervention to evaluate the design (e.g., in Lo & Jeong, 2018; Meyerhöffer & Dreesmann, 2019; Nashaat Sobhy, 2018). So far, there have been just a few CLIL studies not only considering learners as data source but involving them from the start, as for example Banegas (2013), Coyle (2013), or Gupta (2020). In contrast, in the Anglo-American (non-CLIL) context, the concept of student voice and its role for improving local school experiences and educational policies on the basis of participatory research has received considerable attention (Cook-Sather, 2006, 2020; Flutter & Rudduck, 2004; Mitra, 2018). In the context of CLIL, Coyle (2013) argues that learners are indeed capable of contributing to the improvement of their own education while also shedding more light on their learning processes and thus calls for involving students more actively in research, regarding them as "competent social actors" (p. 249). Döring (2020) agrees and maintains, based on her student data, that teachers too could equip students with more agency when planning and preparing lessons.

### **2.3.4 Materials and didactic design in CLIL**

Another area of CLIL research relevant for the present thesis relates to CLIL materials and pedagogical tools and how they are enacted in the classroom. A great part of this research area is concerned with a theoretical elaboration of underlying principles and quality criteria, many of which overlap with quite general current quality criteria for any educational context. The following list of CLIL quality criteria is based on Ball, Kelly and Clegg (2015), Banegas (2017), Mehisto et al. (2009), Meyer (2013), P. Moore and Lorenzo (2015), and Pérez Cañado (2018a). This list contains key aspects that are repeatedly mentioned and is by no means meant to be exhaustive.

- On the social level, CLIL materials should provide opportunities to collaborate in different forms of pair and group work, often through task-based or project-oriented learning, in a safe environment. As such, CLIL materials should be student-centred, which also entails that teachers are seen as mediators and facilitators rather than as "donors of knowledge" (Pérez Cañado, 2018, p. 4). In other words, CLIL materials should offer challenging yet feasible tasks that require the learners to be active, to interact, and to work systematically to reach a certain goal with guidance of their teacher.
- Following constructivist principles, inquiry-based or discovery learning is to be preferred, and critical thinking should be fostered. This requires purposeful didactic planning that involves scaffolding of the learning process in the form of creating manageable steps that progress from the familiar to new concepts, and from lower- to higher-order thinking skills. By guiding the learner through the process, cognitive load is reduced, potentially

enabling students to reach higher thinking skills (including critical thinking) than they would without scaffolding.

- Similarly, linguistic input and output also require scaffolding techniques, creating a language-aware environment. Input scaffolding means to grade, structure, and edit input to enable learners to work through challenging content. This includes creating a sequence of (small) tasks that help learners process the input in a guided way. Moreover, input may be enriched via added visuals, glossaries, explicit explanations, awareness-raising remarks about textual features, or highlighted key points, etc. Likewise, the students' output, i.e., the learners' verbalization when dealing with content, should be supported by staging the process and providing useful phrases, explicit instruction, or model output. Again, this way, learners may be enabled to express cognitive processes, for which they might normally, i.e., without scaffolding, have insufficient linguistic resources.<sup>8</sup>
- Input should be rich and meaningful and may come from authentic sources to boost motivation and make use of the advantages of English as a Lingua Franca. Yet, in the context of CLIL, 'authentic' does not imply that the material is taken from the 'real' world without alteration but that they have not been taken from EFL textbooks for the purpose of language learning and that the tasks the learners are supposed to do are authentic within the subject discipline. As described above, input, especially when it is challenging, which often is the case for authentic texts, requires simplification or adaption to make sure that learners can work with it. Moreover, to ensure variety and to cater for different learner styles, input should come from different types of media and feature various modes, including ICT.
- Given its conceptualization, CLIL materials are said to proffer themselves for cross-curricular, intercultural orientation. As such, CLIL materials should tap into the cultural dimension by adding tasks that foster intercultural communicative competence and/ or consciousness-raising activities that topicalize implied cultural codes, values, and beliefs.
- The materials should promote sustainable learning. This means that the materials should first connect new content to the learners' previous state of knowledge, skills, experiences, or attitudes and secondly make the learning trajectory transparent, including intermediate and final learning goals. Finally, a thorough closing activity should promote potential uptake.
- As for content, Ball et al. (2015) suggest prioritizing three dimensions of content, namely (1) conceptual content, i.e., notions relevant for the discipline, (2) procedural choices, i.e., knowledge enacted, and (3) specific language pertinent for the discursive context. These

---

<sup>8</sup> More information on scaffolding follows in chapter 3.

dimensions may be used as planning tools and their interplay can be exploited to create balanced and rich CLIL lessons.

Despite quite some conceptual work having been done, there is a lack of appropriate CLIL materials and tools that teachers can readily use in their contexts, as already established in section 2.3.3 (Hahn, 2019; Massler, 2012; Meyer et al., 2015; Morton, 2013). Overall, there are not too many ready-made CLIL textbooks, which also results in a paucity of studies directly investigating CLIL materials. Analysing four EFL textbook series that include CLIL-oriented sections, Banegas (2014) found the contents presented oversimplified, an imbalance of language skills featuring too many reading activities, as well as a dominance of lower-order thinking-skill tasks. Another study comparing textbooks was conducted by Maxwell-Reid and Lau (2016) in an EMI setting. Focusing on genre and the construction of technical knowledge in analogical explanation, they demonstrate that two of the three books do not provide enough guidance with regards to text-image relations, pertinent aspects of genre, and constructing technical knowledge. These two books mostly rely on visual representation in the form of diagrams without verbalizing the workings of the systems presented. The authors rightfully point out that understanding and expressing a complex process using appropriate language might not be self-evident to learners and would require more scaffolding, which then the teacher would need to supply. One rare example of a textbook that aims to help students cope with the demands of learning a subject in an FL is *Learning History in English* by Lasagabaster et al. (2021). This book is based on a four-year-long interdisciplinary empirical project and equips history students with linguistic tools tailored to courses offered at the University of the Basque Country. For each module, the authors provide a glossary and practice materials. The practice materials deal with different aspects of historical discourse, including paragraph structure, ways of expressing views and emphasis, graph descriptions, citing and referencing, linking ideas, using passive and nominalisations, giving oral presentations, and more.

In general, however, CLIL textbooks have been criticised for insufficient language support but also for lacking topicalization of cultural dimensions (López-Medina, 2016). CLIL textbooks are often produced for the international market and therefore rarely fit local curricula (Banegas, 2017; Hahn, 2019). The same applies for native-speaker textbooks, reflecting the curricula of the countries of origin, e.g., by focusing on US-American history (Banegas, 2017; Hahn, 2019). Language-wise, using native-speaker materials often entails a mismatch of language level and target age level, e.g., when using native-speaker primary textbooks in secondary education (Hahn, 2019; Morton, 2013). As a consequence, teachers need to regularly create new materials from scratch or rigorously adapt existing materials, which is very time-consuming and can be experienced as arduous (Hahn, 2019; Morton, 2013; Pérez Cañado, 2018). When adapting textual input, P. Moore and Lorenzo (2007) identified three main strategies, namely *simplification*, *elaboration*, and *discursification*. Simplification refers to the reduction of linguistic complexity usually on the sentence-level, resulting in a shortening of the material, and, occasionally in a lack of coherence. Elaboration describes a process of adding phrases to guide the reader through the

text, highlighting what should be remembered. Discursification entails a more global revision of the text, i.e., turning a scientific text into a pedagogic one, which might include adding visuals, glossaries, rhetorical questions, etc., in order to facilitate the learners' comprehension of and engagement with the text. Although not mentioned by the authors, such processes are scaffolding techniques as mentioned previously. In any case, adapting input and preparing scaffolded language is time-consuming but might be worthwhile in order to create materials that are tailored to the learners' needs. As for audio-visual material, Zhyrun (2016) found that students reported positive emotions when watching purposefully created videos. These learners also felt that they could comprehend more when watching 'CLIL videos' rather than YouTube clips. However, finding appropriate audio-visual materials is difficult and creating well-made videos is both laborious as well as expensive. Nonetheless, Zhyrun (2016) argues that the creation and pedagogical preparation of videos or audio input should play a more prominent role in CLIL pedagogy, also considering that audio-visual material has become increasingly popular and practicable for classroom use.

The reported lack of materials has not only been viewed as a burden but, under the right circumstances, could be considered as an opportunity to highlight teacher professionalism. Morton (2013) argues that such views have often been found among immersion researchers and policy-makers, quoting a bilingual project manager: "What you don't need is a textbook [...] What you need are turned-on teachers who are looking at their own kids and can develop resources according to what is needed" (p. 117). The notion of *teachers as designers* implies high degrees of deliberate didactic planning that invites pedagogical innovation and ensures student-centred learning, embodying the foundation of teacher professionalism (see Paniagua & Istance, 2018). Banegas (2017), too, sees creating one's own context-responsive materials as "a personal and professional investment opportunity to reflect on teaching and learning processes" (p. 32). Ball (2018) adds that teachers usually enjoy creating materials and confirms that designing materials can indeed advance teachers' professional development, provided that they are granted enough time to do so, which often is just not the case. Furthermore, Banegas (2017) explains that, especially in the case of CLIL, collaboration between content and language teachers as well as providing adequate teacher training opportunities are prerequisites for teachers to successfully function as designers of their didactic materials. Morton (2013) finds the expectation of teachers as designers demanding both in terms of workload and expertise. Relating to expertise, in a study by Koopman et al. (2014), CLIL content teachers were shown to include language-related actions in their classroom, but their rationales behind these practices seemed to lack theoretical basis. Lo and Jeong (2018) also argue that content teachers often struggle in CLIL because they do not recognize the role of language for their subject and/or do not have the linguistic tools for purposeful scaffolding. However, merely equipping content teachers with language-didactic tools might not be enough to ensure that teachers experience material design as empowering opportunity instead of an added burden. Morton (2013) maintains that this could only work with extensive teacher training that goes beyond general insights of either second language pedagogy or general current pedagogical thinking, but which finds a way to fuse these two perspectives,

reflecting the complexities of CLIL. Good language material might not equal good content material and vice versa, which is why Morton (2013) stresses the importance of approaching material design from an integrative point of view (see also Meyer et al., 2015).

Given that there is a discrepancy between a substantial amount of literature dealing with how CLIL should be and the amount of quality materials and textbooks available, examining how CLIL is pedagogically enacted would be illuminating. However, relatively few studies exist, especially studies based on naturalistic data rather than reported practices (Mahan et al., 2018). Of course, studies on reported practices are also insightful, but they might be more affected by social desirability effects and differences in perception and actual behaviour.

One early example is a study by de Graaff et al. (2007) that identified a number of effective language-related pedagogical practices on the basis of videotaped lessons in triangulation with teacher interviews. These practices include using material that slightly exceed the students' level (both in terms of content and language), facilitating meaning-focused processing (e.g., using worksheets that help learners identify and understand concepts) and form-focused processing (e.g., explaining or correcting relevant linguistic forms), as well as ensuring student output and interaction in the target language, and enabling them to use compensation strategies (e.g., using dictionaries). Considering that subject-specific language might also be unfamiliar in the L1, some of these strategies do not seem to be too different to regular content teaching. In fact, comparing the lesson design of CLIL and non-CLIL lessons, Badertscher and Bieri (2009) found no major differences in pedagogical practice. In a more recent study on self-reported pedagogical practices in the Dutch context by van Kampen et al. (2018), most teachers felt that compared to traditional teaching, CLIL lessons featured oral communication more prominently. Yet, only very few teachers reported that they explicitly dealt with language-related or subject-literacy-related aspects. Much more commonly, in contrast to their regular teaching, these teachers reported that they provided CLIL learners with more creative and diverse input and paid more attention to scaffolding content and language. Some teachers, however, could not identify a difference between the pedagogical practices in CLIL and non-CLIL lessons, with some teachers emphasizing that language was similarly important in regular teaching. Working with naturalistic as well as interview data, Hu and Gao (2020), too, observed very limited attention to linguistic forms and language learning strategies, including language-focused scaffolding, in the context of CLIL science and mathematics.

In contrast to previous studies, Mahan et al. (2018) identified a strong presence of language-related and content-and-language-integrative practices to ease comprehension of input. Directly observing CLIL classroom practice in Norway, Mahan et al. (2018) found consistent use of the L2 (over 80% of the time) and academic language too. In these lessons, academic language was also scaffolded by offering opportunities to negotiate meaning and clarify subject-specific terminology or by using visuals, props, or the L1 to support comprehension. According to the authors, there was also a considerable amount of group work and, especially in science, opportunities to write (lab reports). Looking more closely at content aspects, the learners were presented with clear

content learning aims, while language aims were not shown explicitly. The authors described the explanations of content as rich, lengthy, and accurate and tasks as intellectually challenging, countering fears that CLIL instruction would water down content. In another study in the Norwegian context, Mahan (2020) examined CLIL teachers' scaffolding practices on the basis of naturalistic classroom data. The focus of this study lies on interactional scaffolding, i.e., the ways teachers support their students spontaneously when problems arise. Mahan (2020) found that the CLIL teachers frequently used a range of different scaffolding strategies to facilitate comprehension, including voicing connections to previous knowledge, body language, or, especially in science, supportive videos or animations. Furthermore, teachers considered academic language and subject-specific terminology. As for task-solving scaffolding techniques, Mahan (2020) only observed little evidence of scaffolding concerning the use of strategies or modelling the thinking process. Only scaffolding of uptake was rather consistently present but varied between the three subject teachers. While the social science teacher supported student uptake by rephrasing students' answers and prompting them to elaborate, the geography and science teachers mostly relied on display questions, usually eliciting only short student answers. Although interactional scaffolding implies a reactive nature of teacher practices, the insights presented in this study and the study mentioned previously might also be used to improve CLIL materials proactively. The strategies often used by these teachers indicate areas students struggle with, and therefore it might make sense to incorporate these aspects into CLIL material in anticipation of these issues. Furthermore, the results of these two studies suggest that scaffolding techniques targeting task-completion and cognition could be more prominent in CLIL materials to ensure that learners are well supported not only in comprehension but also in applying knowledge, which might pose a greater challenge than 'just' processing.

### 2.3.5 CLIL and diversity

As discussed in section 2.3.2, CLIL provision has usually been offered on a voluntary basis, often resulting in a so-called "creaming effect" (Rumlich, 2016, p. 89), i.e., creating classes of highly motivated and/ or linguistically gifted students with parental support and greater resources, while the majority of average- and low-performing students remains in mainstream education (see also Broca, 2016; Möller, 2017; Pérez Cañado, 2020; van Mensel et al., 2020). Therefore, CLIL has been criticized for promoting selectivity and elitism (Bruton, 2011, 2013; Dallinger et al., 2018; Paran, 2013). Indeed, a number of outcome-focused studies that factored in initial differences in terms of level of ability (e.g., Artieda et al., 2017; Olsson, 2015; Verspoor et al., 2015), motivation, and/or socio-economic or parental background (Alejo & Piquer-Píriz, 2016; Bruton, 2013; Bulon, 2020; Dallinger et al., 2018; Dios Martínez-Agudo, 2019; Fernández-Fontecha, 2015; Möller, 2017; Pérez Cañado, 2020; van Mensel et al., 2020) found that contextual variables, and especially socio-economic background, seem to account for a substantial part of the learning differences between CLIL and mainstream students. Nonetheless, most of these studies argue that while contextual factors do explain parts of observed learning advantages, so does the type of instruction, i.e., CLIL. While this indicates that CLIL indeed is (potentially) an effective



strategy, it does not counter the elitist connotation of CLIL. In fact, it rather supports it given that these studies all point towards more homogenous and privileged groups of learners. Along these lines, Paran (2013) argues that CLIL is only truly effective in selective contexts with substantial levels of extramural English and general literacy but also where teachers possess high levels of L2 and CLIL competence and students present high levels of general achievement.

Recently, however, CLIL has found its way into mainstream education via whole-school programmes or curricular decisions for entire school types as, for example, in Spain, Italy, the Netherlands (Rumlich, 2020), and Austria (see subchapter 2.2). It is feared that in non-selective settings, (compulsory) CLIL provision could have detrimental effects for learners with low levels of English skills or general academic ability (Broca, 2016; Massler, 2012; Paran, 2013). Earlier studies by Mearns (2012), Fung and Yip (2014), and Mewald (2007) reported that high achievers seemed to benefit more than their low-achieving peers, creating a Matthew effect. Yet, Rumlich (2020) points out that there has hardly been any empirical research on unstreamed CLIL so far. A rare example would be Denman et al. (2018), whose study into attitudinal factors in unstreamed pre-vocational schools indicate positive effects of CLIL within non-selected learners too. However, the authors argue that learners in these contexts would require better support sensitive to different levels of ability. Conducting a literature review about the role of CLIL for migrant learners, Somers (2017) argues that immigrant learners speaking minority languages can reach the same or potentially even higher levels within CLIL programmes compared to mainstream education, as these tend to provide greater pedagogical and linguistic support. Such studies echo with a number of CLIL scholars, who all call for more research into how diverse groups of learners can best be supported. One frequent demand would be better scaffolding and language support but also scaffolding with regards to content, graded to the needs of the learners (e.g., Calderón-Jurado & Garcia, 2018; Lialikhova, 2019; Lo & Jeong, 2018; Roussel et al., 2017). Madrid and Pérez Cañado (2018) further list the following strategies to cater to diverse groups in CLIL settings: flexible, learner-centred teaching methodologies, strategies that allow for differentiation (e.g., adapting tasks for different needs), individualisation (e.g., taking into account information about a learner's academic profile) and personalization (e.g., offering suitable electives and encouraging extramural exposure to the L2 fitting to the learner's preferences) but also team-teaching, ensuring safe learning environments and rapport, addressing multiple intelligences, including ICT, providing ongoing feedback, and clear and slow articulation. Having reviewed pertinent literature regarding CLIL and diversity, Madrid and Pérez Cañado (2018) conclude that while awareness about diversity in CLIL has risen, there is a paucity of research in this regard and few resources for teachers, including materials and opportunities for in-service training (see also Pérez Cañado, 2016a). Similarly, interested in CLIL teachers' views on differentiated instruction, Roiha (2014) found that besides time and physical classroom environment, one of the biggest concerns regards materials. The author further notes that while teachers do differentiate in a variety of ways, they could benefit from a more conscious, deliberate, and systematic approach.

To address the gaps outlined above, the Erasmus+ project ADiBE (CLIL for All: Attention to Diversity in Bilingual Education) set out (1) to investigate if or how CLIL caters for different achievement levels and learner styles across six different countries and (2) to operationalize scientific insights in this regard by creating classroom materials, teacher education modules, and informational videos accordingly (ADiBE, 2021; Pérez Cañado, submitted). As for the first objective, a number of general observations have been reported in Pérez-Cañado (2021): For instance, across the six countries, teachers reported that teaching diverse groups of CLIL learners was challenging and that there indeed was a dearth of appropriate materials considering different needs and strengths of learners. Their main strategy in this regard would be employing learner-centred methods and scaffolding while utilizing peer support, individual attention, and smaller groups were somewhat less common. Making use of mixed-ability groups, multiple intelligences, differing classroom layouts, and ‘newcomer classes’, on the other hand, were rarely exploited. What transpired in several contexts was that teachers tend to cater to individual and different needs in an ad-hoc way, rather than pre-planned, as well as indirectly through various learner-centred methods (see also Bauer-Marschallinger et al., 2021; Nikula et al., accepted). These practices might not always be recognized by learners as pedagogical practices, resulting in rather dissimilar accounts of learners and teachers in the different ADiBE data sets, with the learners usually reporting a more negative picture than the teachers (Pérez Cañado, 2021, see also Bauer-Marschallinger et al., 2021; Siepmann et al., 2021). In the combined data set, learners seem to be appreciative of the teachers’ disciplinary knowledge and language skills but less so of how their teachers deal with translations and scaffold content or how their textbooks and materials consider different needs and preferences (Pérez Cañado, 2021). Moreover, the ADiBE data highlights that in most contexts, diversity of learners does not seem to be adequately reflected in assessment practices.

Another interesting insight pertains to how different stakeholders experience diversity in CLIL settings. For example, Austrian and Finnish teachers noted that CLIL students who opted for the programme are more homogenous than mainstream groups (Bauer-Marschallinger et al., 2021; Nikula et al., accepted). In the Finnish context, participants seemed aware of the selectivity of their programme, which is why upward-differentiation appeared to play a bigger role than inclusive practices. This goes hand in hand with the learners’ assumption that they have to cope on their own, inducing stress but also feelings of pride (Nikula et al., accepted). At the same time, the Finnish team noticed a theme of “equality and ‘the same for all’ principle” (Nikula et al., accepted, p. 1), a phenomenon also found in the Austrian data, where learners argued that they were all equally good at English while stating that everyone possesses different weaknesses and strengths that can complement each other. Their teachers, however, disagreed to different degrees. In the Austrian school where CLIL is optional, the teachers found that there were individual differences but fewer than in mainstream classes, while in the context where CLIL is obligatory, the use of the L2 was believed to widen the gap between differently gifted learners (Bauer-Marschallinger et al., 2021). Interestingly, the UK data indicated a “strong, shared, values-driven understanding of

diversity”, where linguistic and cultural diversity are celebrated and identity-establishing (Coyle et al., 2021, p. 16).

As for the second objective of the ADiBE project, the team has also thought about what diversity means in CLIL and how one could operationalize this for classroom application. On a more conceptual level, Pérez Cañado (submitted) suggests the DIDI framework. Here, *diversity* serves as starting and all-encompassing aspect, including personal traits, individual needs, and differences in terms of cognition, culture, language, learning styles, knowledge, achievement levels, pace, attitudes, experiences, interests, and socioeconomic backgrounds. To account for such a diversity in needs and features of learners, the DIDI framework suggests the use of methods of *differentiation*, understood as catering to different learning needs and potentials present within a group, and *inclusion*, conceptualized as educational model that includes all learners no matter which (special) needs and individual differences they bring to the table, especially if learners are at risk of exclusion or marginalisation (Pérez Cañado, submitted). This should result in *integration*, which is understood as a “consistent response to the diversity of student needs” (Madrid & Pérez Cañado, 2018, p. 245). More practically, the ADiBE team also suggests six principles for designing such materials, namely *teachers as designers*, *dialogic classroom*, *explicitness*, *learner-centredness*, *multimodality*, and *scaffolding*.<sup>9</sup> These principles have been operationalized in a number of CLIL projects, which will be made available online as open educational resource.

Coming back to the critique by Paran (2013), who states that CLIL only works where “learners [...] receive additional language support” (p. 327) and “teachers are educated in CLIL and understand the links between language and content” (p. 329) among other factors, current research indeed tells us that such steps are necessary to help diverse groups of learners cope. However, if diversity is put into the equation and teachers are prepared accordingly, then Paran’s (2013) or Bruton’s (2011, 2013) claim that CLIL could only work in selective contexts does not seem to hold, considering that CLIL has recently found its way into mainstream education and more inclusive contexts and there is some tentative evidence of this being successful (e.g., Denman et al., 2018; Pérez Cañado, 2020; Somers, 2017). Nonetheless – or even because of that – it is the job of researchers, practitioners, and policy-makers to adapt to current educational realities.

### 2.3.6 Linking the strands

In a nutshell, most CLIL-related studies have focused on linguistic outcomes of CLIL, while content-related concerns have only recently attracted notice (Dalton-Puffer, 2018; Fernández-Sanjurjo et al., 2017; Morton, 2020; San Isidro, 2019). Many language-focused studies suggest a positive impact of CLIL, whereas content-centred studies tend to find zero effects (Dalton-Puffer, 2008; Pérez Cañado, 2012; San Isidro, 2019). There are three main issues with this arguably simplified summary. First of all, many (early) outcome-oriented studies did not adequately factor in potential selection biases often found in CLIL practice or did not adequately analyse their data

---

<sup>9</sup> For more information, see the ADiBE website (<https://adibeproject.com/>).

statistically as would be necessitated in comparative studies, and those that do often present less positive results (Bruton, 2011; Graham et al., 2018; Pérez Cañado, 2016b; Piesche et al., 2016). Consequently, the effectiveness of CLIL has been more and more called into question (Pérez Cañado, 2016b; San Isidro, 2019). Secondly, content-focused studies often only test factual knowledge, which might be easier to measure, but this does not sufficiently reflect current competency-based curricula (see, e.g., Gautschi, 2015; Priestley & Biesta, 2017). Finally, looking at the label of CLIL, a split into a language and a content domain appears somewhat absurd. Without language, content cannot be expressed, and without content, language loses meaning.

Interestingly, the issue of lacking integration of content and language learning was already addressed by immersion and ESL research in the North American context in the 1980s (e.g., Snow et al., 1989; Swain & Lapkin, 1989) but found little take-up by the research community at the time. Recently, however, many CLIL researchers have acknowledged and even foregrounded this rationale for both practice and research. For example, Meyer et al. (2015) argue that traditional dichotomies of ‘content vs. language’ (p. 45) should be overcome and that a reconceptualization and better understanding of the relationship of content and language is key for CLIL “to live up to its full potential” (p. 44). In this vein, Nikula et al. (2016) published an entire edited volume dedicated to theorizing integration. Ruiz de Zarobe and Cenoz (2015) even go as far as stating that integration is “the way forward in [...] CLIL for the rest of the twenty-first century” (p. 90).

Before elaborating on the integration of content and language in more detail in chapter 3, the connections of this issue to the other research strands presented in this literature review are discussed in the following. Starting with materials and didactic design, on top of a general lack of CLIL materials requiring teachers to become the designers of their own didactic materials (e.g., Banegas, 2017; Hahn, 2019; Morton, 2013), there appears to be a paucity of content-and-language-integrative resources for teachers to draw from (Meyer et al., 2015; Morton, 2013). Therefore, researchers have called for a more principled approach to didactic design that operationalizes current scientific insights into content and language integration for classroom use. Tying in with research on teachers’ practices and beliefs, (content) teachers often seem insufficiently aware of the connection between content and language, and those that show awareness frequently reported that they struggled with integrating content and language in their didactic designs and classroom practices (Koopman et al., 2014; Lo & Jeong, 2018; Morton, 2013). Overall, it appears that teachers and student-teachers would benefit from a more thorough treatment of content and language integration within pre- and in-service teacher training (Morton, 2013; Pérez Cañado, 2016a).

Turning to research on the students’ perspectives, integration has, so far, only played a minor role, but more research would be warranted. For example, while much of the available literature deals with motivation towards learning an FL, much less is known about their motivation towards CLIL (Somers & Llinares, 2018). Similarly, scant attention has been paid to how learners perceive and understand the integration of content and language learning (Somers & Llinares, 2018). Moreover, participants are usually treated as data source, including studies that intend to create

didactic materials, tools, or interventions. In such studies, especially, more active, participatory roles would be legitimate (Groundwater-Smith & Mockler, 2016). This links back the above-mentioned call for more research-based CLIL materials. Considering the teachers' role, expertise, and knowledge of the inner workings of classroom action, the benefit of having teachers as collaborators for the design of content-and-language-integrative materials seems self-evident (see section 5.1.1 concerning participants roles). Yet, learners, too, can contribute to creating a successful design (Cook-Sather, 2006, 2020; Coyle, 2013). By giving learners a voice in the development process, these materials might better reflect the learners' needs and wishes, which would ultimately facilitate the target audience's acceptance thereof.

Recent discussions on diversity in CLIL also link to the design of content-and-language-integrative materials. To ensure that CLIL does not only cater for gifted and privileged students but promotes learners of all ability levels and educational needs, it is necessary to create materials that foster subject-specific literacy skills and provide ample explicit language support and scaffolding, taking individual differences into account (Calderón-Jurado & Garcia, 2018; Lo & Jeong, 2018; Madrid & Pérez Cañado, 2018; Paran, 2013; Roiha, 2014; Roussel et al., 2017; Somers & Llinares, 2018). This, in turn, requires better teacher support in that regard, raising teachers' awareness and equipping them with practical tools (Madrid & Pérez Cañado, 2018; Morton, 2013; Pérez Cañado, 2016a).

To summarize, the integration of content and language and its operationalization connect the different strands of research presented in this literature review, namely outcome-oriented studies, participants' perspectives, materials and didactic practices, and diversity in CLIL. Overall, it seems that the current main question in CLIL research is not *if*, or for which aspects of learning, CLIL might be (most) effective but, in light of the widespread practice of the approach, how one can make it work *better* for a variety of learners. Integrating linguistic and content-related perspectives and operationalizing the theoretical outcomes of this discussion seem to be key to this endeavour. The following chapter delves into the topic of content and language integration by outlining the foundations and different approaches to integration and discussing their applicability for classroom use.

### 3. Conceptualizing the integration of content and language

*“A word devoid of thought is a dead thing [...] but thought that fails to realize in words remains a stygian shadow.” (Vygotsky, 1987, p. 255)*

The relation between thought and language lies at the core of integrating content and language learning. Given that cognition is not directly observable, it remains unclear to what extent language can adequately represent cognitive processes or whether language mediates or even expands thought processes (Dalton-Puffer, 2013). Overall, most cognitive scientists would corroborate such a tight relation (Dalton-Puffer, 2013) as would a variety of language theories, such as the Sapir-Whorf hypothesis or Fodor’s (1995) *Language of Thought Hypothesis* and a number of related approaches (Heine, 2010). Although such a theoretical discussion would be interesting, it is not central to the experienced realities of most learners. As the quote above illustrates, language and thought become meaningful through each other, i.e., in a social context, and this is especially relevant in the educational sphere, where language is the main tool for sharing cognitive processes and functions as “primary evidence for learning” (Mohan et al., 2010, p. 221). As Schleppegrell (2001) famously put forward, “[i]t is through language that school subjects are taught and through language that students’ understanding of concepts is displayed and evaluated in school contexts” (p. 1). As such, the relation of cognition, content learning, and language is clearly important for learners in any educational setting, resonating with the often quoted notion that “every teacher is a language teacher” (van der Walt & Ruiters, 2012, p. 85) already introduced on the very first page of this thesis. The interplay of these aspects gains even more importance when the language of instruction is not the learners’ first language.

In the case of CLIL, it appears that, for a long time, it was assumed that by being ‘dual-focused’, content and language learning would automatically be integrated, and the only question would be how to ‘balance’ content and language, both in research and practice (see chapter 2). For example, Lyster (2007, 2017) argues for a counter-balance approach to integrate content and language learning, meaning that in content-focused settings (hard CLIL), learners’ attention should be proactively and reactively guided towards linguistic aspects via form-focused tasks, and in a language-based setting (soft CLIL), content concerns should be highlighted more prominently. Tedick and Lyster (2019) maintain that shifting between content and language foci increases depth of cognitive processing. To operationalize this shift, Tedick and Lyster (2019) presented their CAPA model, which can be understood as a blueprint sequence defining four central stages for integrating content and language. Tedick and Lyster (2019) explain that first, one should provide meaningful contexts for the target feature (*contextualization*). Then, the teacher should help learners take note of the feature in focus and guide them towards discovering patterns in the text provided (*awareness*). The next phase is called *practice*, indicating that learners should be provided with opportunities to employ the target feature in a controlled but still meaningful way. Here, Tedick and Lyster (2019) add that corrective, form-focused feedback is beneficial. Finally, the last step, *autonomy*, is intended to increase the learners’ fluency, confidence, and motivation by encouraging them to make use of the feature flexibly and autonomously. As such, the model by

Tedick and Lyster (2019) starts with a focus on content, then moves to a focus on language and lastly shifts back to a content focus. Overall, this model reflects its origins in applied linguistics and echoes the PPP approach (*presentation, practice, production*) often featured in ELT course books (Harmer, 2015), adapted for focusing on form in CBI contexts. However, it should be kept in mind that form-focused approaches in content-based settings might not be always accepted by content teachers or learners. For example, a teacher involved in an intervention study by Tedick and Young (2018) reported that “a focus on form during content instruction ‘hijacked’ his teaching” while “integrating functional language was fun” (p. 316), suggesting that everything that is not closely connected to the content would feel wrong in two-way-immersion. Thus, Tedick and Young (2018) argue that “[t]eachers must consider how to attend to language without sacrificing sense-making around the subject area content” (p. 315). In other words, an organically integrated approach is needed, i.e., an approach that goes beyond combining content and language. This appears to be especially important in contexts where the instructional focus is on content learning (hard CLIL).

From a theoretical point of view, after treating the conceptual fusion of content and language “like a hot potato” (Dalton-Puffer et al., 2010, p. 288) for a long time, considerable headway has been made recently (Dalton-Puffer et al., 2018; de Graaff, 2016; Donato, 2016). However, as Donato (2016) put it, “[w]hat is lacking is conceptual clarity and a cohesive pedagogy [...] about what it means to design, implement, and carry out a program that purports content and language integration” (p. 29). Meyer et al. (2015), too, argue that research insights in this regard have not been made viable for actual teaching settings, resulting in CLIL not “liv[ing] up to its full potential” (p. 44).

This chapter explores different theoretical approaches that organically integrate content and language learning and discusses their viability for classroom practice in the context of history education. The chapter starts with two theories that are strongly meaning-based, namely sociocultural theory (SCT, subchapter 3.1), followed by systemic functional linguistics (SFL, subchapter 3.2). Then, the 4Cs framework and its elaboration, the pluriliteracies approach, are presented as constructs that combine aspects of SCT and SFL to fuse content and language learning (0). Finally, the concept of cognitive discourse functions (CDFs), another notion assumed to conceptualize and operationalize an organic integration of content and language learning, is outlined and discussed in subchapter 3.4.

### **3.1 Sociocultural theory**

In terms of a wider theoretical embedding, for some, the way to conceptualize integration is via sociocultural theory (SCT). In the early-20<sup>th</sup>-century Soviet Union, psychologist Lev Vygotsky developed a theory postulating that (1) the relationship of thought and language is dynamic and reciprocal, thus thought and language are inextricably linked (Vygotsky, 1987), and (2) that all learning takes place in a sociocultural context, highlighting the importance of interaction for cognitive development (Vygotsky, 1978, see also Cammarata et al., 2016; Lantolf et al., 2018).

Though incomplete due to his early death in 1934, Vygotsky's works and elaborations by his colleagues and students were eagerly taken up and advanced in the context of educational and linguistic research in the late 20<sup>th</sup> century (Lantolf et al., 2018).

According to Vygotsky (1987), "[w]ords and other signs are those means that direct our mental operations, control their course, and channel them toward the solution of the problem confronting us" (pp. 106–107). Put differently, humans require semiotic tools to shape and control their thinking processes. As Moate (2010) points out, especially in educational settings, our primary tool for meaning-making is language, as it is the "tool of engagement between learner and teacher, learner with subject, learner with learner" (p. 41). Following Vygotsky (1987), this tool has two functions, a social one and an intellectual one. The intellectual function is now also known as *linguaging*, which refers to the process of using language to shape knowledge and make meaning (Swain et al., 2015). Young children, Vygotsky (1987) explains, tend to speak to themselves when thinking (i.e., *Private Speech*) before internalizing this process, resulting in *Inner Speech*. As such, a primarily social process (*intermental*, e.g., talking) develops into a psychological one (*intramental*, e.g., thinking) (see Swain et al., 2015).

The emphasis on the interrelations between thought and language are not the only aspect of SCT providing a frame for conceptualizing the integration of language and content learning. The core element of SCT is the understanding that learning is a social, interactive activity in sociocultural context (Vygotsky, 1978). Moate (2010) explains that in SCT, "[t]he social dimension is more than a safe, supportive environment: it is the area within which learning actually occurs" (p. 39), making it markedly different from individualised theories of learning. From a SCT perspective, interaction with others where learners verbalize their cognition in the social sphere, bouncing off ideas and re-shaping them, is necessary to advance cognitive development (Swain et al., 2015). This is in line with Moate's (2010) argument that in an ideal CLIL classroom, exploratory talk is given enough room to make both language and subject learning manageable and productive. Donato (2016) further points out that this means that learners are socialized into "the discursive practices of an academic content area" (p. 32) rather than being the mere recipients of input. For this process of socialization, which some also call *acculturation*, experts need to mediate their disciplinary ways of thinking and expressing, e.g., talking like a historian (see Donato, 2016; Mahan et al., 2018; Moate, 2010). Conceptually, such an understanding of learning dismisses any possibility of splitting language from content learning. Moreover, according to Donato (2016), research in the CBI context has shown that learning becomes more effective if teachers are aware of this connection and consider it in their teaching by topicalizing disciplinary language in context rather than teaching content and language disjointly, e.g., in the form of pre-teaching linguistic structures before engaging with content.

In terms of manageability, another core concept of SCT is the zone of proximal development, abbreviated as ZPD, which describes the difference between what a learner can do on their own and what they could do with support (Swain et al., 2015). Although similar to Krashen's (1987) well-known *i+1* concept, i.e., that input should be comprehensible at a level that exceeds the



learners' level by one stage, Swain et al. (2015) explain that Krashen's model only considers the level of language, while the ZPD encompasses all dimensions of activity. Moreover, Vygotsky did not conceptualize learning in stages, waiting for advancement through the various levels. Instead, he called for creating situations in which development is encouraged (Swain et al., 2015). In other words, the ZPD is understood as development potential, where educators, *experts* in SCT terminology, create opportunities for the learners, the *novices*, to progress (Holzman, 2018). The concept of scaffolding (see 2.3.4), though not used as a term by Vygotsky, strongly relates to the ZPD and as Swain et al. (2015) put it, "[s]caffold seems a helpful verb to operationalize the meaning of a ZPD" (p. 25).

As a means to integrate content and language learning, Tedick and Lyster (2019) have outlined helpful ways of scaffolding student production and comprehension. They differentiate between *verbal scaffolding*, *procedural scaffolding*, and *instructional scaffolding*. Verbal scaffolding for production, according to Tedick and Lyster (2019), includes different types of corrective feedback, display questions, and referential questions. For verbal comprehension, this type of scaffolding makes use of redundancy, meaning that messages are repeated in different ways, e.g., by using synonyms, paraphrases, or examples, by adding paralinguistic features, and/ or by moderating the volume, intonation, and speed of speech. Procedural scaffolding for production aims at creating opportunities in which the learners use the target language cooperatively (e.g., think-pair-share set-ups, dyads, cooperative learnings groups, peer learning, or peer feedback) and actively (e.g., role-plays, simulations, presentations, or debates). Here, the authors note that teachers need to structure the tasks well and provide scaffolds for the language to be used. For procedural scaffolding aimed at comprehension, Tedick and Lyster (2019) recommend making use of routines and activity frames in order to increase predictability and to give clear instructions. Instructional scaffolding for production, Tedick and Lyster (2019) state, involves explicit instruction and modelling necessary language, such as chunks or useful phrases. Instructional scaffolding for comprehension, on the other hand, aims at enabling learners to understand the content, language, and instructions by using various devices such as graphic organizers, props, graphs, imagery, maps, multi-media input, or interactive whiteboards.

According to Holzman (2018), most empirical research looking into shared activity within the ZPD has focused on the expert-novice relationship. Yet, Holzman (2018) strongly argues for paying more attention to peer-to-peer interaction and collaboration, adding that Vygotsky himself did not restrict learning to a dyadic process, highlighting that "[l]earning awakens [...] in cooperation with peers" (Vygotsky, 1978, p. 90). Looking into oppositional talk and argumentation in CLIL social sciences classrooms, Hüttner and Smit (2018) observed two main patterns, which both highlight the importance of peer interaction for disciplinary skills. Firstly, there was joint construction of subject-specific language and content ("learning-focused argumentation", p. 297) and secondly, these were then enacted in interaction ("expertise-focused argumentation", p. 297). The role of the teacher in this study was that of the expert, making meta-comments as well as modelling and scaffolding argumentation (Hüttner & Smit, 2018). Interested in peer-to-peer

scaffolding within CLIL, Lialikhova (2019) compared group interactions of 9<sup>th</sup>-grade Norwegian CLIL learners, grouped according to proficiency levels. She found that both the mid- and high-achieving group scaffolded each other's historical learning within their ZPD, e.g., by drawing on previous knowledge or providing corrective feedback, ultimately eliciting *higher-order-thinking skills (HOTS)*. The group consisting of low achievers, however, collectively avoided HOTS and did not collaborate to the same extent as their class mates, requiring the teacher to intervene frequently. Lialikhova (2019) concludes that teachers need to provide a higher degree of scaffolding for low-achieving students. This ties in with the point made by Donato (2016) that students present different ZPDs and therefore would require differentiated ways of assistance, as any form of mediation would be ineffective outside the ZPD. Lialikhova (2019) further maintains that opting for heterogenous rather than homogenous groups might not be helpful for low-achieving students as weaker students might feel silenced or marginalized when their high-achieving peers take up the role of experts. To be able to use group-based learning in mixed-ability contexts without enforcing a deficit model, deliberate planning might be key. When planning for heterogenous groups, tasks need to be designed in a way that not only enables but also requires weaker learners to contribute something the stronger students can benefit from, potentially exploiting everyone's strengths and avoiding situations where high achievers keep explaining to their low-achieving peers (see Tomlinson, 2001). For such a set-up, it might also make sense to split learners into homogenous groups first and adapt the degree of scaffolding accordingly, before mixing them together again, for instance, for an information gap activity<sup>10</sup> (see, for instance, the *CLIL pages* in Kilbey et al., 2018).

Having briefly reviewed core themes of SCT, looking at learning from a sociocultural perspective seems appropriate in the context of CLIL. Indeed, SCT has functioned as a basis for the conceptualization of learning within CLIL settings (Banegas, 2013; Coyle et al., 2010; Lialikhova, 2019; Moate, 2010). According to Dalton-Puffer et al. (2010), "sociocultural theory furnishes the base-line understandings in which learning in CLIL classrooms can best be understood and how it should consequently be viewed" (p. 8). As a result, a considerable number of CLIL studies, especially those investigating co-construction of meaning or the integration of content and language learning on the level of classroom talk, embedded their studies within a sociocultural perspective of learning (e.g., Barwell, 2016; Dalton-Puffer, 2007; Devos, 2015; Evnitskaya & Morton, 2011; Heimes, 2011; Hüttner & Smit, 2018; Lialikhova, 2019; Mahan et al., 2018; Mahan, 2020; Morton & Jakonen, 2016; Nikula, 2010). Heimes (2011), for instance, investigated history learners' theories about the relations of content and language learning from a sociocultural and psycholinguistic perspective. He found that conceptually and from the learners' perspective, a CLIL approach enables sociocultural language learning processes because adding an FL enriches and supports the development of history skills, ultimately facilitating the entry to potential vocational or academic communities. However, he also identified a great need for supporting

---

<sup>10</sup> An *information gap activity* is a classroom activity where learners are given different pieces of information which they then need to share to successfully complete a task.

these processes by integrating content and language learning more consciously, e.g., by stressing the role of language for subject-specific skills.

Turning to the operationalization of content and language integration from a sociocultural perspective, it seems that there is little research that translates SCT systematically and tangibly for classroom use within specific subjects. To fill this gap, researchers such as Morton and Llinares (2017) or Donato (2016) recommend complementing the sociocultural perspective with other theories such as systemic functional linguistics (see subchapter 3.2). Overall, rather than a means for operationalization, it appears that SCT alone is more useful as a framework to understand learning in CLIL, which includes the link between language and disciplinary ways of thinking and expressing. In this vein, Donato (2016) calls for providing (CBI) teachers with a deep understanding of sociocultural principles to support them against the challenge of content and language integration. While such an understanding would definitely help educators to conceptualize content and language integration generally, it does not provide them with a systematic pedagogy or clear guidance for implementation. Nonetheless, some practical implications have been put forward, many of which in connection to the ZPD and scaffolding (see, e.g., Banegas, 2013; Donato, 2016; Lialikhova, 2019; Mahan, 2020). For instance, Donato (2016) argues that scaffolding should dynamically consider the learners' ZPD. Furthermore, he advises against "frontloading" (p. 29) scaffolding, i.e., pre-teaching useful vocabulary or phrases, as this disconnects language from the context of the discipline. Another theme in SCT-based pedagogical implications for the integration of content and language learning relates to the social dimension and the role of classroom talk. Moate (2010), for example, suggests transforming the traditional initiation-response-feedback (IRF) pattern of classroom interaction into initiation-discussion-response-feedback (IDRF) to allow for more exploratory talk and meaningful, student-centred interaction. In this regard, Donato (2016) recommends including information gap activities that require meaningful sharing and discussion of academic content rather than exchanging phrases or pieces of factual knowledge.

### **3.2 Systemic functional linguistics**

Systemic function linguistics (SFL) has been regarded as a suitable complementation to SCT to provide a more tangible framework for content and language integration in the context of BE (Donato, 2016; Llinares et al., 2012; Morton & Llinares, 2017; Walker, 2010).

After a general introduction to key terms of SFL, this subchapter zooms in on SFL-based approaches integrating content and language in the domain of history. First, central features of historical discourse as identified from an SFL perspective are outlined and connected to studies investigating CLIL learners' realization of these features. Subsequently, approaches in accordance with this line of research that were developed for classroom use are discussed.

### 3.2.1 Key concepts and overview

Systemic functional linguistics (SFL) is a meaning-based theory of language originally devised by MAK Halliday (Halliday, 1975, 1993; Halliday & Matthiessen, 2014), which has had huge influence on educational linguistics research in a wide area of contexts (Coffin, 2006; Morton & Llinares, 2017). As the name suggests, SFL understands language as *systemic* entity because users of language can choose from an array of various lexicogrammatical options, i.e., different language patterns that could go together, in order to express and construct meaning (Coffin, 2006; Halliday & Matthiessen, 2014). For Halliday (1993), “learning is learning to mean, and to expand one’s meaning potential” (p. 93), increasing the number of options available and the purposefulness and appropriateness of these choices in different instances and domains (see also Coffin, 2006). Such a systemic approach to language further implies that “functionality is intrinsic to language”, meaning that “the entire architecture of language is arranged along functional lines” (Halliday & Matthiessen, 2014, p. 31). Not to be confused with communicative functions of language, Halliday identified three metafunctions of language (Halliday & Matthiessen, 2014):

- The ideational function is used for construing our experience and making sense of the world (“language as reflection”, Halliday & Matthiessen, 2014, p. 30).
- The interpersonal function relates to the interaction with others, establishing and maintaining relationships (“language as action”, Halliday & Matthiessen, 2014, p. 30).
- The textual function enables the other two metafunctions by organizing “the discursive flow and creating cohesion and continuity as it moves along” (Halliday & Matthiessen, 2014, p. 31).

Another important pillar within SFL theory is the role of the social context and the view of language as a social-semiotic tool (Halliday & Hasan, 1989). In other words, the three main meanings, i.e., three metafunctions, are dependent on the situational and cultural context (Halliday & Hasan, 1989). According to Halliday and Hasan (1989), three dimensions of the situational context affect the way language is used, also known as register variables (see also Christie & Derewianka, 2008; Llinares et al., 2012; Rose & Martin, 2012):

- *Field* describes the type of activity participants are engaged in or the topic of the interaction. As such, it relates to the ideational metafunction.
- *Tenor* refers to the different roles and relationships between the participants and applies to the use of the interpersonal metafunction.
- *Mode* relates to the channel of the interaction, i.e., written or spoken language, affecting the textual metafunction of language.

On the level of cultural context, the process of meaning-making depends on the genres that are being enacted, which are understood as “social processes for achieving purposes within the culture” (Christie & Derewianka, 2008, p. 7, see also Rose & Martin, 2012). According to Rose and Martin (2012), this process follows a number of stages in order to reach certain goals, and these steps vary from one culture or discipline to another.

This all amounts to a contextualized, socially embedded theory of language, which according to Coffin (2017) “provides powerful tools for the systematic and rigorous analysis of how meanings are made through language” (p. 92). Similar to SCT, SFL provides a framework that organically

integrates content and language, as these two domains cannot be uncoupled. Bartlett (2017) therefore, considers SFL as “clearly socioculturally oriented” (p. 376) and views SCT as “one of the most significant and enduring theoretical influences within the SFL model” (p. 377). It should be kept in mind, however, that SFL “is primarily a theory of language, rather than a theory of learning” (Coffin, 2017, p. 91), which lies in contrast to SCT but which might also be the reason why these two theories could complement each other well. In fact, SFL has also been described as a language-based theory of learning and knowledge (Zydatið, 2007).

Considering how closely intertwined language, content, and learning are from an SFL perspective, research into bilingual programmes often uses an SFL-based framework, both as an analytical tool but also as a way to operationalize the integration of content and language learning in bilingual classrooms (Coffin, 2017). As for the former, there is quite a considerable body of research regarding subject-specific texts and expected discourse practices (e.g., Achugar & Schleppegrell, 2005; Christie & Derewianka, 2008; Coffin, 2006; Schleppegrell, 2004). These elaborations have convincingly demonstrated that the language of education differs not only from our everyday language but also from discipline to discipline. As such, subject-specific language does not have any ‘native speakers’, and thus it needs to be taught in school. Another line of research investigates how (CLIL) learners realize subject-specific language (e.g., Dalton-Puffer & Llinares, 2015; Järvinen, 2010; Llinares & Morton, 2010; Llinares & Nikula, 2016; Morton, 2010) or how these learners progress in developing disciplinary linguistic resources and meaning-making strategies (e.g., Llinares & Pascual Peña, 2015; McCabe & Whittaker, 2017; Morton & Llinares, 2018), elucidating the process of learning subject-specific language in an L2. These lines of research are discussed in section 3.2.2, focusing on typical features of historical discourse as identified from an SFL perspective and how CLIL students enact historical literacy. Concerning the operationalization of integrating content and language learning, several approaches within an SFL framework have been developed, most of which work with notions of *genre*, *teaching/learning cycles*, or *R2L (reading to learn)*. These approaches are briefly summarized in section 3.2.3

### 3.2.2 Subject literacy: the case of history

Traditionally, the concept of literacy refers to the learners’ ability to read and write, but recently, this term has been extended to procedural knowledge and the ability to express oneself within different subject areas (Pavón Vázquez, 2018), including history. As such, literacy plays a central role in the development of discipline-specific competences. Llinares et al. (2012) use the term *subject literacy* to cover *genre*, i.e., the various subject-specific text types, and *register*, i.e., the lexicogrammatical resources, both in the written and oral mode. Such an understanding entails that subject-specific language skills go beyond knowing and appropriately using field-specific terminology, as it includes the adequate construction of meaning and the fulfilling of different functions within a certain domain. In other words, working on subject literacy entails developing control of subject-specific genres, which can greatly differ from discipline to discipline (McCabe & Whittaker, 2017; Morton, 2020). While some general features of academic language have been

recognized across different languages and academic disciplines, their (culturally and linguistically appropriate) realization might differ (Achugar et al., 2007).

### 3.2.2.1 History genres

In the context of the subject history, Coffin (2006) and Christie and Derewianka (2008), among others, have comprehensively identified typical features and requirements of a range of text types for history as a discipline. An overview of different types of history genres is presented in Table 1, which is based on Christie and Derewianka (2008), Coffin (2006), and Llinares and Pascual Peña (2015):

Table 1. Main genres in history (based on Christie & Derewianka, 2008; Coffin, 2006; Llinares & Pascual Peña, 2015)

overall function	genre and subtypes		specification
describing (non-chronological) genres	1	period study	describing a historical period
	2	site study	describing a historical site
recording genres	3	autobiographical recount	retelling major events of a historical figure's life
	4	biographical recount	
	5	historical recount	retelling historical events chronologically
	6	historical account	retelling and explaining the chronological sequence of historical events
explaining (non-chronological) genres	7	historical explanation	(multi-layered) explanation of ...
		a. factorial explanation	a. why a historical event occurred, focusing on different factors contributing to the outcome
		b. consequential explanation	b. the consequences of certain actions or events
arguing (rhetorical) genres	8	historical argument	
		a. exposition	arguing why a certain historical interpretation is valid or significant
		b. discussion	gauging different perspectives before positioning oneself
		c. challenge	countering someone else's interpretation (e.g., a historian's view)

Given their different functions, these genres require different cognitive operations, which in turn are expressed with different linguistic resources, both of which with increasing complexity (Llinares & Pascual Peña, 2015). In non-chronological description, evaluative language or establishing nexus to other historical aspects are typically not expected (Christie & Derewianka, 2008). Reporting genres, on the other hand, rely on chronological organization of texts, whereas in explaining genres, grammatical metaphors are expected to provide a stringent and concise network of factors, causes, or consequences (Christie & Derewianka, 2008; Coffin, 2006). Finally, arguing genres consist of a number of moves and require more variation in tenor to engage in evaluation (Christie & Derewianka, 2008; Coffin, 2006).

### 3.2.2.2 Features of historical language

Looking at linguistic features of historical discourse more generally, the following have been identified:

Historical texts are usually packed with information, reflected in dense and complex language, making it difficult for learners to decipher the meanings (de Oliveira, 2010; Donato, 2016; Lorenzo, 2017; Schall-Leckrone & McQuillan, 2012; Schleppegrell et al., 2004). There are several reasons contributing to this high density of information, closely linked to the function of historical texts. To begin with, complex webs of cause-and-effect relations are central elements of historical discourse. Usually, there is no singular event causing a historical development but rather a number of factors influencing one another, aggregating to a certain outcome (Achugar & Schleppegrell, 2005; de Oliveira, 2010). For instance, the subject of the sentence “the French Revolution exhilarated political change all over the world” contains a great number of concrete events and developments (as does the object of the sentence). Moreover, sometimes, singular agents might be replaced with abstract concepts or nominal groups because identifying them is impossible, deemed unnecessary, or simply does not fit the predominant historical discourse (Achugar & Schleppegrell, 2005; de Oliveira, 2010). This way, actors are separated from their actions (Achugar & Schleppegrell, 2005; de Oliveira, 2010). For example, rather than saying “when Nazi-Germany invaded Czechoslovakia, the British and French finally realized that appeasing Hitler was not possible”, history textbooks would formulate this differently, as found in the GSCE textbook *Modern World History*: “The fall of Czechoslovakia, however, had convinced the British and the French that appeasement had failed” (Walsh, 2002, p. 82). The use of nouns rather than verbs to create dense and usually abstract and semantically complex sentences is also known as grammatical metaphor (Järvinen, 2010; Ryshina-Pankova, 2016). Having good control over grammatical metaphors has been considered a key element of academic success (Lorenzo, 2017; Morton, 2010). Lorenzo (2017), who qualitatively investigated historical literacy skills of upper secondary bilingual learners, found that CLIL learners are indeed capable of using nominalisations in historical discourse. Taking a longitudinal approach, Whittaker et al. (2011) found that, over the course of four years, CLIL learners considerably improved their use of nominal phrases in written historical discourse. Investigating the use of grammatical metaphors by lower secondary Finnish CLIL learners and international students (i.e., all instruction in English, English = L1 or L2), Järvinen (2010) found that the CLIL learners’ writing in English contained more grammatical metaphors and was thus more intricate and lexically dense than their writing in Finnish. Interestingly, the international students in her study obtained a considerably higher score compared to the CLIL students, irrespective of the language the CLIL learners used.

A nominalised style does not only allow presenting events as *things*, but it also structures the process of reasoning (Achugar & Schleppegrell, 2005; de Oliveira, 2010). By using nominalised groups, causality is often established within a clause rather than between clauses, e.g., using structures like “led to”, or “exhilarate” and “convince” in the examples above instead of subordinate clauses linked via cohesive devices (Achugar & Schleppegrell, 2005; Schleppegrell et

al., 2004). Comparing written and spoken productions of CLIL and non-CLIL history learners, Llinares and Whittaker (2010) found that the CLIL students expressed causality more often between clauses than within in contrast to their L1-instructed peers. While both groups overused “and” to link their ideas, the L1 group’s writing was much more distinct from their oral use of language, suggesting that the CLIL group especially struggled with written expression of causality in the context of history.

The use of non-finite dependent clauses, like “[i]gnoring the Cherokee’s treaty rights, ...” or “[t]o silence further criticism, ...” (Achugar & Schleppegrell, 2005, p. 306), is another typical way of expressing causality within clauses. However, in such asyndetic structures, causality is not expressed explicitly, making it difficult for learners to grasp cause-and-effect relationships (Donato, 2016; Lorenzo & Dalton-Puffer, 2016). In terms of finite subordinate clauses expressing causality, Achugar and Schleppegrell (2005) found that history textbooks often use cohesives that imply contrast or concession (“however”, “despite”, ...) or linking devices that conflate time and cause (e.g., “after”, “as”) rather than cohesive devices typical for cause-effect relations (e.g., “because”, “so”; see also Coffin, 2006). In general, it seems that historical discourse expresses causality in a variety of ways, and many of these are not extremely obvious, making it hard for learners to understand the meaning of historical texts (Achugar & Schleppegrell, 2005). Looking at language production, a study by Llinares and Morton (2017) suggests that expressing causality is indeed difficult for CLIL learners. Their functional analysis of interviews and role-plays demonstrates that CLIL learners only rarely use prolonged moves enhancing cause-and-effect relations and overall seem limited in their expression of causality. For example, prototypical linking devices like “because” or “so” make up the majority of cause-and-effect markings, especially in the role-plays. In an earlier related study, Llinares and Morton (2010) compared CLIL learners’ explanations in classroom discussions and individual interviews. While the students demonstrated their ability to produce longer explanations using a range of different appropriate lexicogrammatical features in the interviews, classroom discussions did not offer the same opportunities as the teacher would quickly evaluate a student’s contribution, “adding it to the ‘official’ explanation”, thereby ending the learner’s turn (Llinares & Morton, 2010, p. 61). Looking at historical writing of upper secondary CLIL learners, Lorenzo (2017) observed instances of sophisticated, multifactorial explanations and asyndetism, suggesting that at later stages and in written mode, CLIL learners are indeed capable of expressing causality in a domain-appropriate way. In a longitudinal analysis of the same corpus with a focus on formal aspects, Lorenzo et al. (2019) could show that over time, these learners’ historical writing became more sophisticated also in terms of structural metrics, like syntactic pattern density, diversity, or complexity.

Another central feature of historical discourse relates to the interpersonal function, namely using language of appraisal to be able to react to and evaluate past events (Coffin, 2006, see also Martin & White, 2005). For example, using modal verbs (e.g., “may”, “might”) or modal adjuncts (e.g., “maybe”, “potentially”) enables the historian to mark that other interpretations of a historical source, event, or causal relations are possible and negotiable (Coffin, 2006). In this regard,



Lorenzo (2017) reports that the CLIL learners in his study rarely evaluated or took a stance, but when they did, their opinions came across in a subjective, overt manner, leaving little room for other views. Moreover, the evaluations identified by Lorenzo (2017) were often “accomplished with unsubstantiated opinions and without any real analysis or perhaps understanding of its purpose and effect” (p. 37).

Coming back to typical features of historical discourse relating to evaluative language, historians usually use intensifiers (e.g., “large”, “significantly”, “very”, “true”) or hedges (e.g., “small”, “slightly”, “somewhat”, “of sorts”) to signal the force or focus of an evaluation (Coffin, 2006, see also Martin & White, 2005). The use of attitudinal lexis is another central element in expressing appraisal in history (Coffin, 2006). In relation to historical genres, these different features have been clustered to describe various *voices* typical in historical discourse, namely the *recorder voice* and *appraiser voice*, which can be further split into *interpreter* and *adjudicator voice* (see Coffin, 2006, for more information). In a longitudinal study into the development of appraisal strategies and historical voices of Spanish CLIL students, McCabe and Whittaker (2017) could show that the learners became more appropriate and attuned to the prompted genres in their use of historical voice over the course of four years. Especially the text by the higher-rated group exhibited various strategies for opening up the dialogue and allowing different interpretations. Focusing on four students over the course of four years, Morton and Llinares (2018) observed that students with better language skills (as rated by the teacher) used a greater variety of linguistic features for the purpose of appraisal, thus developing appropriate voice, whereas lower-rated students struggled in this regard. In terms of development, in the first three years, a sharp increase in the frequency of using resources from the appraisal framework was reported, while in their year-four texts, these numbers dropped again, which Morton and Llinares (2018) explain with the prompt of the task, which triggered different genres than was intended in some cases. In terms of task or activity type, Dalton-Puffer and Llinares (2015) demonstrated that role-play and interview elicited the most instances of evaluative language, followed by group work, presentation, and whole-class discussion.

Another frequently mentioned feature of historical discourse is *backshifting*, i.e., moving back and forth in time as well as linking different past, current, and parallel events, which requires good control of tenses (Lorenzo, 2017; Lorenzo & Dalton-Puffer, 2016). Moreover, as different perspectives should be integrated in such shifts, indirect speech needs to be mastered. (Lorenzo & Dalton-Puffer, 2016). Backshifting is necessary to create multifactorial historical narratives that do the complex past justice but which also allow taking a personal stance, ultimately leading to the development of historical voice (Lorenzo & Dalton-Puffer, 2016). Coffin (2006) adds that recording genres require the learner to identify central questions, events, people, or periods and to relate these to one another. This necessitates language for ordering and structuring but also for expressing cause-and-effect (Coffin, 2006).

### 3.2.3 SFL-based approaches to integrate content and language learning in bilingual history education

As mentioned in chapter 2, CLIL/ CBI teachers often do not really consider themselves language teachers and, apart from vocabulary teaching, usually take an implicit approach to teaching (discipline-specific) language (Dalton-Puffer, 2007; Hüttner et al., 2013; Kong et al., 2011; Lo & Jeong, 2018; Morton, 2019; Skinnari & Bovellan, 2016; Tan, 2011). Consequently, it is not surprising that Morton (2010) could not observe evidence of genre-related teaching in his case study. Llinares and Pascual Peña (2015), too, report that history CLIL teachers focus on facts in their oral questions, irrespective of the genre targeted by the prompt. According to the authors, this might have to do with the common assumption that fact-oriented teaching would be cognitively and linguistically more accessible and would also relate more to the traditions of the subject. However, Llinares and Pascual Peña (2015) argue that enabling teachers to scaffold different genres orally via teacher training programmes might be key to empowering students to actively participate in subject discourse.

In fact, several researchers have argued that scaffolding the various steps and realizations of subject-specific genres would be an effective way to integrate content and language teaching (Llinares et al., 2012; Meyer et al., 2015). One impactful approach to genre-based teaching has been developed by the so-called *Sidney School* of SFL (e.g., Martin & Rose, 2008; Rose & Martin, 2012; Rothery, 1994). This approach has been developed via a series of literacy-focused large-scale action research projects over more than three decades and combines Halliday's SFL theory with Bernstein's sociological theory of pedagogic discourse (Rose, 2009, 2014). Extending the sociocultural concept of socialisation (see also Bartlett, 2017), Bernstein (1999, 2003) argued that pupils need to be socialised into codes of power, i.e., academic discourse, to ensure that educationally alienated learners can succeed as well. To enable the learners to move from *horizontal discourse*, i.e., everyday uses of language, to *vertical discourses*, i.e., specialized, academic language, Bernstein (1999) called for making explicit the rules and linguistic structures that shape specialized knowledge, competences, and literacies. Along these lines, exponents of the Sidney School ventured into identifying and mapping the genres of schooling, making accessible the inner workings, i.e., patterns of stages and their features, of various genres for pedagogical use, including the development of explicit metalanguage to be used in class (Rose, 2009, e.g., Martin & Rose, 2008; Rose & Martin, 2012). In this vein, Rothery (1994) created a concrete writing pedagogy, termed *Teaching/Learning Cycle* (TLC), consisting of three steps. The first step is called *deconstruction* and refers to the process of guiding the students through the deconstruction of a representative text of a certain genre, including its cultural context, social purpose, stages, and central linguistic features. Next, students and teachers would jointly construct another text of the same genre (*joint construction*) before learners try to create their own example of this genre (*individual construction*) (Rose, 2009; Rothery, 1994). For the stage of joint construction, Macnaught et al. (2013) argue that teachers should build knowledge via *semantic waves*, moving back and forth between abstract, specialized meanings and simpler, tangible meanings, i.e.,

unpacking and repacking technicality. Conducting an intervention study centring on the concept of semantic waves, the authors report that working with this concept helped teachers become aware of the role of language for their subject. In an exploratory study focusing on semantic waves, Lo et al. (2020) observed that teachers often tend to unpack technicality but neglect the repacking. The authors argue that equipping teachers with tools to create these semantic waves could improve content-and-language-integrative education.

An example for a study operationalizing content and language integration via the TLC in the context of CLIL history education was provided by Lo and Jeong (2018), who conducted a collaborative genre-based intervention study. Two student groups at grade 8 at a secondary school with an English-medium programme in Hong Kong participated, one being considered to be an average group and the other one was reported to be an 'elite' group. After a written essay on ancient Greco-Roman civilisation as pre-test, the teacher guided the learners through the deconstruction of a historical exposition text on the topic of the renaissance in line with the TLC approach. In the joint construction phase, the students collaboratively reflected on their pre-test essay and collected ideas for a new essay in connection to the renaissance, which they later wrote down on their own (individual construction). These texts served as post-intervention essays. Based on analytic rubrics, the examination of the pre- and post essays indicates that these learners improved their performance both in terms of content and language, with the weaker group making somewhat more headway. Especially scores for organisation and systematicity increased, as most students were able to apply the structure of the target genre after the intervention. However, it should be kept in mind that the second time around, the students had collaboratively prepared for the individual construction unlike in the pre-test, where they were only given some guiding questions in the prompt. Overall, the teacher involved in this project found the approach effective, particularly for low-achieving students, and predominately in terms of language. While being convinced of the advantages of the approach, this teacher also expressed concerns that content teachers might struggle with this approach and "may need to be 'psychologically' and 'practically' prepared before implementing genre-based pedagogy" (Lo & Jeong, 2018, p. 43). From the learners' perspective, it was reported that they perceived some improvement themselves and that they appreciated the integrated character of the new approach, highlighting, for example, the usefulness of model texts or the connectives.

Another closely related approach to the operationalization of content and language integration is the R2L methodology, *Reading to Learn* (Rose & Martin, 2012), which essentially constitutes an elaboration of the TLC (Rose, 2018). Here, the practice of reading takes on a more central role since pedagogically prepared and purposeful reading activities form the starting point of a pedagogical cycle (Rose, 2018).<sup>11</sup> This pedagogy formed the basis of an international action

---

<sup>11</sup> First, teachers prepare learners to read texts that may be beyond their current skills in a step-by-step manner, followed by detailed and guided reading. Then, teacher and learners engage in 'sentence making', re-organizing and rewording key sentences of the reading passage, which might be followed by spelling activities. Subsequently, the learners should construct their own sentences before engaging in joint rewriting of passages. Finally, learners are guided to construct complete texts of the genre in question (joint

research project called TeLe4ELE, *Teacher Learning for European Literacy Education*, which was targeted at improving literacy skills in various subjects, including history (Coffin et al., 2013; ELINET, 2015). Overall, the learners' reading and writing skills improved while teachers reported appreciation of the integrated approach to literacy (Whittaker & García Parejo, 2018).

Rose (2018) stressed that the R2L approach is not only a classroom methodology but also a professional learning programme. In general, it seems that SFL-based approaches operationalize the integration of content and language learning through comprehensive teacher training programmes. In the context of history education, *Building Academic Literacy through History* represents an example of such a programme (Schleppegrell et al., 2008). Under this label, a range of summer schools, academic programmes, and online courses have been offered since 2003, enabling teachers to guide learners through the deconstruction and sense-making of history texts (Schleppegrell et al., 2008). Reporting on a case study, Schleppegrell et al. (2008) present evidence that students benefit from this approach, producing well-structured essays and presentations of historical reasoning, which also reflects in better results at standardized history exams.

In collaboration with two history teachers participating in a summer course focused on a functional approach to historical literacy, Achugar and Carpenter (2012) conducted a design experiment in a setting with both English native speakers and English language learners in the US. The intervention of this study, spanning over one term and consisting of three focal lessons, aimed at enabling learners to understand historical documents via text analysis and metalinguistic explanation about how meaning is constructed in such texts, i.e., 'reading like a historian'<sup>12</sup>. Testing their reading comprehension of historical documents before and after the intervention, Achugar and Carpenter (2012) report improved results for all learners. When looking more closely at the written responses, they found that all learners produced more intricate and dense answers, with more instances of technical terms and a greater variety of clause types after the intervention. In terms of expressing interpersonal meanings, the learners started to take a more academic position by distancing the author from the reader and guiding readers better through their texts.

Another, more extensive design-based project was conducted by J. Moore et al. (2018) in collaboration with teachers and literacy coaches from six different schools. This three-year project involved the design of an SFL-based approach aiming at supporting English language learners in developing a metalanguage that helps them read, write, and engage with subject-specific texts. Following a needs analysis on the basis of interviews with literacy coaches and analyses of curricular texts, pedagogical materials and professional development materials were designed, and the participating teachers took part in these teacher training workshops. For the purpose of evaluation, classrooms were observed, student performances were analysed, and teachers were interviewed. Typically for DBR, this process happened in iterative cycles, allowing the design and

---

construction). For more information, see Rose (2018) on R2L in general or Whittaker (2018) for R2L in CLIL subjects.

<sup>12</sup> For more information on the intervention, see Carpenter et al. (2014).

theoretical propositions to evolve through the identification and analysis of critical events of the process and their impacts on the design. Zooming in on the DBR processes in the context of the subject English language arts, the authors concluded that a DBR approach helped them make complex theories like SFL “usable” (J. Moore et al., 2018, p. 1044). Moreover, based on their data, they argued that the use of appraisal metalanguage helped learners identify and interpret explicit and implicit attitudes, but in the process, the authors also realized that the instructional meaning got lost if lexical items were treated in an isolated or narrow way. Looking at the students’ written production, J. Moore et al. (2018) observed that learners tended to just restate the claims of a text, creating texts devoid of any real analysis. The authors realized that the initial teacher development materials did not adequately prepare the teachers for making clear and accessible what it means to analyse in the respective discipline, resulting in teachers only providing examples of analysis without giving explicit guidance and direction. In relation to this, the authors came to the conclusion that

it is not just the linguistic knowledge itself, or the understanding of SFL concepts, that is important for teachers and students to develop. Instead, teachers need to understand why they are using the metalanguage, how it can be used to talk about meaning, and how to teach it in ways that support students to achieve curricular goals. (J. Moore et al., 2018, p. 1035)

In other words, it does not suffice to help teachers understand linguistic concepts. Rather, understanding the rationale behind certain decisions and ways to translate this knowledge into classroom applications are key.

In the two studies mentioned above, the researchers closely worked with the teachers; in the case of J. Moore et al. (2018), extensive teacher training was provided, and in the study by Achugar and Carpenter (2012), the researcher frequently met up with teachers explaining concepts, providing feedback, and discussing plans. However, without intensive guidance, teachers often struggle with SFL-based approaches, as SFL is deeply rooted in linguistics (J. Moore et al., 2018; Schall-Leckrone & McQuillan, 2012). In the context of a two-year action research project aiming at developing a teacher training module focused on teaching historical literacy, Schall-Leckrone and McQuillan (2012) analysed the perceived preparedness of student-teachers to teach history to English learners in US secondary education. In the first research cycle, the student-teachers’ awareness of the role of language for history teaching increased, but they did not feel more prepared to teach historical literacy. Consequently, in the second cycle, Schall-Leckrone and McQuillan (2012) focused more on the practical side of scaffolding, language-focused activities, as well as modelling language objectives, resulting in a statistically significant improvement of perceived preparedness. However, the authors also note that the student-teachers would still need more practice before applying SFL-concepts in their own teaching, and some history teachers did not demonstrate sufficient knowledge of linguistic structures to an extent where they could guide English learners in their analysis of historical texts. Moreover, some future history teachers felt unsure whether this approach might be confusing for learners and/ or too time-consuming (Schall-Leckrone & McQuillan, 2012). Another concern reported relates to the terminology of SFL and whether one could use these terms with learners, and if not, how teachers could break them

down for their students. For these reasons, Schall-Leckrone and McQuillan (2012) call for strategic cooperation between content and language education specialists to prepare future teachers to gain better understanding of the linguistic demands of different subjects and to viably proceduralize this knowledge for classroom implementation.

Along these lines, Schall-Leckrone and Barron (2018) developed SFL-based history materials together with a teacher and a student-teacher within a mentoring framework. The didactic materials designed and implemented focused on reading and writing historical explanations and took the form of the TLC outlined above. Before this study, the experienced teacher, who got in contact with SFL during her studies in a seminar offered by the researcher, could not really make sense of the theory:

[W]hen you were teaching us about SFL [in the history methods course], I was like, “What are we talking about?” I kept ... hearing it and ... still [didn’t] get it. (Schall-Leckrone & Barron, 2018, p. 217)

After the intervention, however, Schall-Leckrone and Barron (2018) reported that the teacher felt more comfortable using the approach and also perceived student improvement. Using an SFL-based analytical tool, the authors also found that the students’ writing scores increased throughout the project. The student-teacher too, reportedly, felt that participating in this project helped her contextualize her coursework, resulting in coherent teacher preparation. Thus, Schall-Leckrone and Barron (2018) argue that classroom-based mentoring, first between researcher and the content teacher and later between the teacher and student-teacher, ensured that both experienced and future content teacher developed the tools necessary to implement genre-based teaching.

To summarize, SFL-based approaches can be effective in linking content and language learning in content subjects. However, SFL is a complex theory of language whose classroom application is not straightforward. Therefore, teachers require extensive support, either in the form of practice-related teacher training seminars or, even more effectively, in the form of mentoring relationships. Therefore, this approach only seems to resonate with teachers who are willing and able to put in time and effort. Moreover, most of these studies do not consider how students experience this approach and what they might wish for. As argued in subsection 2.3.3.3, ultimately, the students’ voices also play an important role in the usefulness and success of a new approach.

Finally, it should be kept in mind that SFL-based studies, for the most part, focus on reading and writing texts. While this is an important area of subject literacy, which also greatly affects academic success (considering that most testing happens in written form), it does not present the full picture of what learners need to be able to do. Nor does an approach focused on reading and writing accurately reflect what is going on in many (CLIL) history classrooms, especially in the European context, which are predominately oral (Dalton-Puffer, 2007, 2011). Some studies have acknowledged that the oral perspective has been somewhat avoided and have offered insights into how meaning is expressed in oral classroom settings (e.g., Dalton-Puffer & Llinares, 2015; Llinares & Morton, 2010, 2017; Llinares & Pascual Peña, 2015). These studies tend to argue that

rather than focusing on whole, self-contained genres, it would make sense to focus on individual stages.

### 3.3 From the 4Cs framework to a pluriliteracies approach

Turning to CLIL-specific approaches to content and language integration, one early holistic conceptualization was provided by Coyle et al. (2010), known as the 4Cs framework (Figure 1). Arguing against the metaphor of the 'language bath', this framework calls for a strategic and purposeful integration of four components of CLIL, namely *content*, *cognition*, *communication* (which is understood as a synonym for language), and *culture*, set within a specific and variable context. To be more precise, content learning is

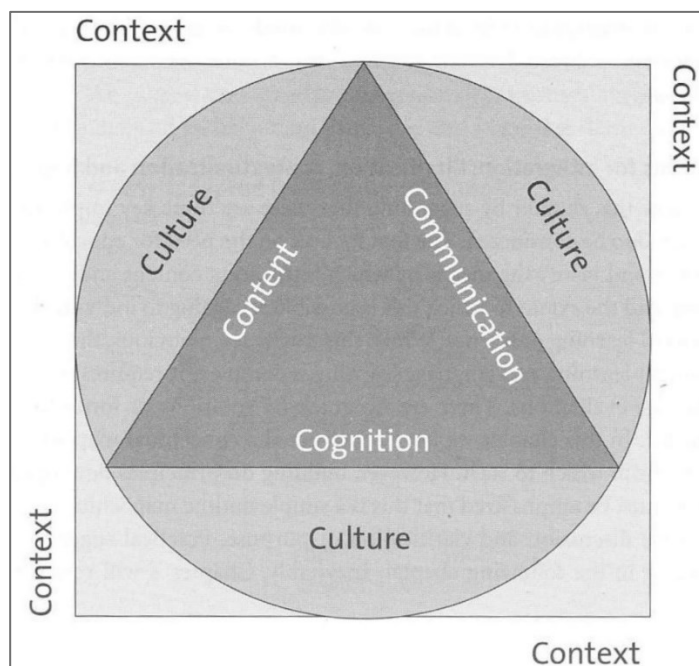


Figure 1. 4Cs model (Coyle et al., 2010, p. 41)

not restricted to learning declarative knowledge but should also entail “creating their own knowledge and understanding and developing skills” (Coyle et al., 2010, p. 42). For these purposes, a range of different thinking processes need to be activated.

To ensure that learners are up to the task, the linguistic dimension and demands of both content and cognition need to be analysed and subsequently made accessible and tangible for the learners. Moreover, following a sociocultural notion of learning, these processes require interaction to promote learning. As a conceptualization for the links between language, content, and cognition, Coyle et al. (2010) offer the so-called *language triptych*, highlighting three different perspectives of language in CLIL, namely *language of learning*, *language for learning*, and *language through learning*. *Language of learning* deals with the language that students require to attain basic concepts and skills for the content subject. For example, understanding past tense is a necessity for history. *Language for learning* is concerned with skills that are necessary for the students to participate in the CLIL learning setting, e.g., debating, asking questions, memorizing, etc. *Language through learning* focuses on the fact that learning is much more effective if there is deep processing and active involvement. This should be the case when content and language learning are integrated because students actively have to articulate their understanding. Concerning the relation between cognition and communication, Coyle et al. (2010) state that in CLIL, there is a risk of mismatching linguistic and cognitive level, either restricting the possibilities of learning or overwhelming learners. As a solution, the authors suggest strategic and purposeful planning using the *CLIL matrix*, which is an adapted version of a model by Cummins (1984), moving from low

linguistic and low cognitive demands to low linguistic and high cognitive demands, and then to low cognitive - high linguistic demands, before finally reaching a level where both cognition and communication are challenging. Turning to the role of culture, Coyle et al. (2010) maintain that our cultural background determines our interpretations of the world, linking culture, communication, and content. Working and interacting in a language different to one's own opens up opportunities to approach content differently and would thus not only foster intercultural awareness but also deep learning. This is especially important in the subject history, where multiperspectivity and understanding others are central skill sets (see section 4.2.3 for more information).

While this framework shows teachers which areas in CLIL are relevant and conceptually interact, it does not straightforwardly translate into classroom practice (Meyer et al., 2015). As Coyle et al. (2018) argue, “[i]t captures the ‘what’ rather than the ‘how’ of CLIL” (p. 354). To fill this gap and to offer a more comprehensive approach to content and language integration, Meyer et al. (2015) developed the concept of *pluriliteracies*, called *Graz Group model* (Figure 2):

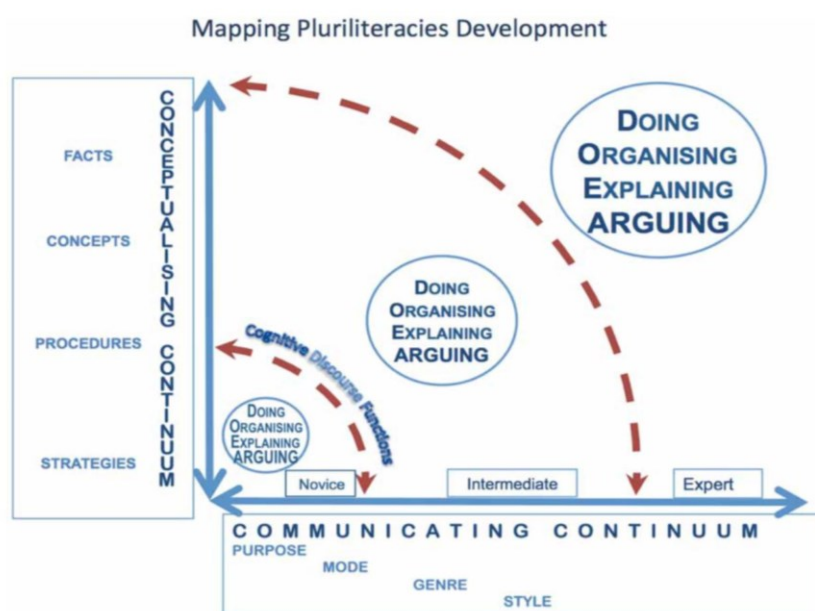


Figure 2. Mapping pluriliteracies development (Meyer et al., 2015, p. 49)

This model is aimed at making teachers aware of the interactions of content and language while also functioning as an “idealized pathway into a discipline” (Meyer & Coyle, 2017, p. 201) and a guide for evolving pedagogical practice. In other words, this model intends to visualize how learners become more and more capable of expressing subject-specific conceptual knowledge and skills in a subject-specific way on various levels of language in a plurilingual setting such as CLIL. To move from subject novices to subject experts, learners need to work on the links between the *conceptualizing continuum* and the *communicating continuum*, i.e., *doing*, *organizing*, *explaining*, and *arguing* science/history/geography and their corresponding genres with the help of their teachers (Meyer & Coyle, 2017). As such, this model combines a sociocultural view of learning with a functional approach to language encapsulated in the notion of genre and subject literacies.



According to Meyer et al. (2015), the pluriliteracies approach presents an extension of the 4Cs in the following way:

*C-Content* in and by itself is meaningless unless it is conceptualised. To actively construct knowledge and to promote subject-specific literacies, learners need to conceptualise content in ways that are appropriate to the subject *C-Culture* [...] it is this subject *C-Culture* that determines how the *C-Cognition* is put to use in the way that *C-Content* will be conceptualised and how the *C-Communication* is used to (co-)construct knowledge. (p. 51)

Like the 4Cs framework, this model argues that linking content and language does not suffice to ensure deep processing and effective learning. Instead, Meyer and Coyle (2017) argue that deeper learning, which they define as “successful internalization of conceptual content knowledge and the automatization of subject specific procedures, skills and strategies” (p. 199), depends on the students’ acquisition of subject-specific literacies. However, the Graz Group does not limit their understanding of subject-specific literacies to large-scale concepts such as genre and register usually associated with SFL. Instead, they argue that another smaller unit might work well as building blocks for more complex genres and, ultimately, as gateways for progressing along the two continua (Meyer et al., 2015; Meyer & Coyle, 2017). These ‘micro-genres’ are called *Cognitive Discourse Functions* (CDFs) and are discussed in the following subchapter.

Apart from the conceptualization of content and language integration, the rationale of the Graz research group was to operationalize scientific insights into the interrelations of content and language learning for CLIL teachers and teacher educators. Therefore, the pluriliteracies team set up a practically-oriented repository for teachers, sharing conceptual information in various forms (documents, videos, FAQ section, presentations, journal articles, etc.) as well as practical resources and sample materials, available online (ECML, 2020).

### **3.4 A construct of cognitive discourse functions (CDFs)**

#### **3.4.1 Rationale and aims of the CDF construct**

As argued in the previous subchapters, to operationalize a content-and-language-integrative approach, a notion is needed that provides “a zone of convergence between content and language pedagogies” (Dalton-Puffer, 2013, p. 216) which is accessible for subject teachers too. As Dalton-Puffer et al. (2018) maintain, “approaches to language-aware subject teaching that are exclusively anchored in the world of linguistics and language education are in danger of being experienced as transgressive or even meaningless by content-subject educators” (p. 7) and, thus, would have little chance of transforming classroom practice.

To address this need, Dalton-Puffer (2013) proposed *a construct of cognitive discourse functions* (CDFs). CDFs are defined as “verbal routines that have arisen in answer to the recurring demands while dealing with curricular content, knowledge items, and abstract thought” (Dalton-Puffer, 2016, p. 29). In other words, CDFs are language patterns humans recurrently use to express and share their thought processes, e.g., when we are explaining, categorizing, or hypothesizing, creating “observable analogues” of cognition (Dalton-Puffer, 2013, p. 220). In the educational context, such processes often form the basis of learning objectives in today’s competency-based

curricula, as exemplified by the following learning aims defined in the Austrian curriculum for secondary commercial colleges:

*The students can*

- *explain the functioning of the Austrian democracy and compare these with other models*
- *describe different forms of rule and leadership structures and discuss their effects on states and the society*
- *name peace-making measures for political stabilization, assess the importance of international organizations to secure peace and evaluate their actions in regard to sustainability*

(Austrian Federal Ministry for Education, 2014, pp. 91–92, official translation)

These subject-specific actions need to be ‘language’d during lessons and in testing situations, which often is not intuitive for learners (Nashaat Sobhy, 2018). As such, CDFs highlight how language is “a natural concern of non-language educators [...], commensurate with the educational goals they want to reach in their respective subjects [...], formulated in terms that are accessible to them from within the ambit of their own disciplines” (Dalton-Puffer & Bauer-Marschallinger, 2019, p. 33). In other words, CDFs make the linguistic demands of subject-specific learning objectives more visible and relevant for content teachers, who often assume that the linguistic dimension would not fall into their area of responsibility (see also Evnitskaya, 2019; Morton, 2020). Along these lines, the concept of CDFs is intended to present a “conceptual base for language-aware pedagogical planning and pedagogical action that speaks to subject educators in terms they can accept as ‘theirs’ and thus has a chance of being translated into practice” (Dalton-Puffer & Bauer-Marschallinger, 2019, pp. 32–33). Ultimately, the value of any approach to the integration of content and language learning lies in the practitioners’ acceptance and readiness to take up said approach, and the CDF construct seems to consider classroom viability and (content) teacher needs.

This focus on practicability is reflected in the formal extent of Dalton-Puffer’s (2013) construct, featuring only seven discourse functions central for academic discourse. At this point, it should be noted that the concept of academic or cognitive discourse function is not new. Starting with the seminal work by Bloom (1956), various scholars from different disciplines have taken up the concept of academic or cognitive language functions, which resulted in a multitude of different frameworks with a variety of labels that are similar but different nonetheless (see more in the following section). Hence, another aim of Dalton-Puffer’s (2013) construct was to create a heuristic of central academic language functions that is manageable yet comprehensive, systematizing and condensing previous research on academic language functions.

Finally, working with complete genres might not be a suitable approach to integrate content and language learning in all situations or contexts, especially where classroom practices heavily rely on spoken interaction. Here, a more modular approach consisting of smaller elements as offered by the concept of CDFs could come in useful. In other words, CDFs present a finer granularity than genres (Dalton-Puffer, 2013), which might or might not be combined into more extensive, subject-

specific texts, allowing the teacher to adapt more flexibly to the learners' needs and contextual factors (Meyer & Coyle, 2017; Nashaat Sobhy, 2018).

### 3.4.2 Theoretical background

In order to create a construct able to bridge subject and language pedagogies, Dalton-Puffer's (2013) CDF construct is equally anchored in educational and linguistic theory. Starting with education, cognitive learning objectives have been playing a central role in educational research in the past decades, dating back to Bloom's (1956) highly influential taxonomy of thinking skills (Dalton-Puffer, 2013). Bloom's taxonomy was intended as a tool to develop curricula and test designs by creating a shared language about learning objectives which go beyond factual knowledge as traditionally emphasized. This taxonomy identified six types of cognitive skills relevant for classroom practice, which are *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation* (Bloom et al., 1956). These dimensions are sorted in ascending order in terms of assumed complexity, resulting in pyramidal visualisation of thinking skills well-known to many educators around the world (Dalton-Puffer, 2013). Recently, this hierarchal order has been criticized for implying an "essentialist ranking" of thinking skills, indicating that, for example, *evaluation* would always be more complex than *synthesis*, or that *knowledge* or *comprehension* would always be inferior to *application* (Dalton-Puffer, 2013, p. 221). To break up this highly hierarchical character of the taxonomy while also stressing the dynamic nature of cognitive skills, Bloom's taxonomy was revised by L. W. Anderson and Krathwohl (2001). The revised version is two-dimensional, presenting a knowledge dimension (*factual*, *conceptual*, *procedural*, and *meta-cognitive knowledge*) and a cognitive process dimension (*remember*, *understand*, *apply*, *analyse*, *evaluate*, and *create*<sup>13</sup>). L. W. Anderson and Krathwohl (2001) argue that this way, educators might become more aware about the relations between different knowledge types and cognitive processes, enabling them to formulate more precise learning aims. While the hierarchical order is still present in this framework, the overall structure is more flexible, resulting in a less pronounced hierarchy (Krathwohl, 2002). Moreover, the process domain was labelled with verbs rather than nouns to emphasize the active nature of these cognitive processes (Krathwohl, 2002).

The work by Bloom, L. W. Anderson, Krathwohl and colleagues has been the starting point for a number of other constructs that map thinking skills and learning aims. For example, Biggs and Tang (2011) present a hierarchically structured framework of verbs to be used for the formulation of learning outcomes in tertiary education, which ultimately should help lecturers to align learning objectives, teaching methods, and assessment. Their framework begins with quantitative processes, i.e., processes that increase knowledge, such as IDENTIFY (level 1: *unistructural*) or DESCRIBE, LIST, or ENUMERATE (level 2: *multistructural*), before moving up to processes that deepen knowledge (qualitative phase), like EXPLAIN, ANALYSE, APPLY (level 3: *relational*) or THEORIZE, GENERALIZE, or HYPOTHESIZE on the highest level (level 4: *extended abstract*) (Biggs & Tang, 2011).

---

<sup>13</sup> *Synthesis* was renamed *create* and swapped places with *evaluate* (Krathwohl, 2002).

While the frameworks mentioned thus far clearly focus on the cognitive dimension of learning, a number of scholars have also elaborated on the linguistic demands of these thinking skills and academic language skills. These often draw upon Cummins' (1980) dichotomous conceptualization of language skills into *basic interpersonal communicative skills (BICS)* and *cognitive/ academic language proficiency (CALP)*. According to Cummins (2008), "BICS refers to conversational fluency while CALP refers to students' ability to understand and express, in both oral and written modes, concepts and ideas that are relevant to success in school" (p. 71). Typically, interpersonal language is embedded in rich context, whereas academic language is context-reduced and cognitively demanding (Cummins, 1980, 2017). Cummins' distinction of BICS and CALP emerged as a reaction to the frequently reported finding that immersion learners would quickly become conversationally fluent, while certain areas of formal linguistic proficiency and overall educational success would lag behind, especially compared to their L1 counterparts (Cummins, 1980; Evnitskaya, 2019). Cummins (1980) adds that BICS and CALP can empirically be distinguished in first and second languages and argues that the L1 and L2 CALP are closely interrelated, meaning that L2 CALP improves more quickly if L1 CALP is developed further.

While many scholars agree with Cummins that educational success highly depends on academic language skills (e.g., Bailey & Butler, 2003; Dalton-Puffer, 2013; Evnitskaya, 2019; Kidd, 1996; Schleppegrell, 2004; Thürmann, 2010), the binary nature of the construct, with its implied deficit approach, often resulting in oversimplified misrepresentations of the model, have been criticized, as noted by Cummins (2008, 2017) himself (see also Dalton-Puffer, 2013; Meyer et al., 2015). Although he stands by his differentiation of BICS and CALP, Cummins (2017) explains that "literacies are multiple, contextually specific, and constantly evolving" (p. 68), arguing for a more multidimensional continuum while also maintaining the usefulness of distinguishing between conversational fluency and academic language skills in certain contexts. One variable here is the degree of context-embeddedness as well as cognitive demand, which, according to Dalton-Puffer (2013), "puts CALP in connection with notions of thinking skills" (p. 226).

An example for combining academic language proficiency and curricular cognitive demands was provided by the CRESST research group (e.g., Bailey et al., 2007; Bailey & Butler, 2003; Butler et al., 2004), who developed frameworks of academic language for the application in the US K-12 school system. Based on an analysis of national and state content standards, TESOL standards, linguistic demands of standardized testing, the language of textbooks and classroom discourse, as well as teachers' linguistic expectations, Bailey and Butler's (2003) framework explores the language demands in content subjects. Taking a functional approach to language, they found a range of language functions learners are expected to use for the demonstration of content knowledge. In the classroom-related data, Bailey and Butler (2003) observed that explanation, description, and comparison were central elements, amongst others. In the curricular documents, the type of functions varied extensively, but they were usually encapsulated in verbs prompting certain expected linguistic behaviours, such as JUSTIFY, DEFINE, CONTRAST, or ENGAGE, among many more. The broadly based CRESST framework was then applied and operationalized for more

specific contexts (e.g., Butler et al., 2004, in mathematics and science; Bailey et al., 2007, on academic reading in various content areas).

In Europe, too, a similar project called *Language in Other Subjects* was brought into being by the Council of Europe (CoE), investigating the linguistic demands of a number of different subjects such as sciences, mathematics, geography, or history. It is aimed at developers of curricula, teaching materials, and tests, but also at teacher trainers and teachers “of subjects sometimes quite wrongly described as ‘non-linguistic’” (Beacco et al., 2010, p. 5) to highlight that language forms an essential part of subject-specific skills beyond knowing technical vocabulary. Amongst other language-related features of academic language, they identified a number of discourse functions, which they then structured according to Mohan’s (1986) knowledge framework, which is an early but influential construct conceptualizing the link between content and language. In the case of history education, Beacco (2010) compiled inventories and descriptions of strategic, discursive, and formal competences needed for teaching of history, taking into account the social components of history education, subject-specific competences, and general educational values within history didactics. As part of *formal competence*, Beacco (2010), lists over 25 cognitive operations, such as ANALYSE, INFER, ILLUSTRATE, CLASSIFY, DESCRIBE, COMPARE, EXPLAIN, etc. He adds that for each of these cognitive operations, one could specify the linguistic demands for their production in reference to the Common European Framework of Reference (CEFR), which, however, might vary from one discourse type to another. Indeed, the CoE commissioned a project aimed at the development of CEFR-linked descriptors for mathematics and history/ civics based on the discourse functions identified in the *Language in Other Subjects* project (Moe et al., 2015). These descriptors were developed via multiple feedback cycles with the help of over 350 experts in the field, including teachers, teacher trainers, researchers, and CEFR specialists. Overall, the experts and teachers participating in this study assumed that 12/13-year-old learners of mathematics and history would require level B1, whereas 15/16-year-old students would already need B2 to succeed in these subjects. These subject-specific demands were then mapped onto descriptors for A2 to B2, divided into the different language skills and organized according to discourse functions, as can be seen in Table 2 on the next page. Working with over 25 different discourse functions, the categories and descriptors provided by Moe et al. (2015) show substantial overlap and are not clearly defined. For example, EXPLAIN/ speaking/ B2 mostly describes an evaluation process and does not seem distinctly different from EXPRESS OPINIONS/ DISCUSS. EVALUATE/INTERPRET, on the other hand, seems to put evaluating and hypothesizing into the same category. So, while it can be argued that this framework highlights the importance of language in content subjects from the angle of different language skills, it does not do so in a neat and useable way.

Table 2. Sample descriptors for speaking in history/ civics and mathematics (Moe et al., 2015, pp. 69–70)

	A2	B1	B2
EXPLAIN	Can explain how to do something or what has been done in simple sentences	Can explain and give reasons for why things, related to history/civics or mathematics, are the way they are, and why something is a problem in a straightforward way	Can give the advantages and disadvantages of various solutions and options. Can explain different phenomena, (for instance, historical or mathematical processes), results or views on topical issues clearly
EXPRESS OPINIONS, DISCUSS	Can say, in a simple way, what s/he thinks about something, or whether s/ he is for or against something	Can explain why s/he is for or against something in a straightforward way	Can argue for her/his points of view and discuss the pros and cons of opposing positions or ways of solving a task in detail. Can discuss and explain her/his attitude towards a topical issue and make hypotheses. Can develop a clear coherent argument, linking ideas logically and expanding and supporting his/her points with appropriate examples
EVALUATE/ INTERPRET	Can state whether something is good or bad, positive or negative in simple sentences	Can give some reasons for why a source is reliable, or why something is an advantage or a problem	Can evaluate different sources or ideas and solutions to a problem. Can make hypotheses about causes, consequences and hypothetical situations

In summary, many educational researchers seem to agree that working with academic language functions is a reasonable way to specify the linguistic demands in content education, thereby connecting content and language pedagogies. However, there is a wide array of different functions. To be more precise, after reviewing fifteen different frameworks, Dalton-Puffer (2013) identified 57 academic language function (see also Lackner, 2012). At the same time, there seems to exist no clear conceptualization what certain functions entail, leading to considerable overlap and idiosyncrasies, which ultimately makes them problematic to work with in educational contexts. As Morton (2020) argues

the use of these verbs can be quite messy, with sometimes different verbs referring to the same thinking skill, or the same verb being used to describe different thinking skills. This can lead to teachers and teaching materials giving misleading information to students about the tasks they have to do. (p. 10)

This might not only result in non-transparent pedagogical design, it can also be problematic for assessing learners (Morton, 2020). Ahern et al. (2018), too, argue that inconsistent use of labels and metalanguage results in inequality, as only those able to “intuitively work out what teachers expect” (p. 27) or those that are being helped by parents or tutors would know what to do in tests or home assignments.

To create a shared basis of labels and increase the practicability of academic language functions, Dalton-Puffer (2013) set out to systematize and condense previous constructs into a manageable number of prototypes. For this process, Dalton-Puffer (2013) drew on functional theories of

language. As mentioned before, Dalton-Puffer (2013) argues that while SFL-based notions such as genre and register grasp communicative routines and intentions, they are “too large” (p. 229) for oral classroom settings and too focused on writing and reading. Instead, she proposes speech act theory as linguistic basis (see Austin, 1962; Searle, 1969). Assuming that by using language humans perform social acts, this theory centres on communicative intentions and the performative nature of language, i.e., *do[ing] things with words* (Austin, 1962). As such, this theory focuses on the speaker rather than communicative *interaction*. What is more, speech act theory has, so far, mostly been applied to analyse interpersonal and everyday uses of language, such as apologizing or promising, but in school, too, learners are expected to verbalize communicative intentions (Dalton-Puffer, 2013). In contrast to everyday uses of language, the focus in educational settings lies on how learners express illocutionary acts (i.e., the underlying intentions) about dealing with and sharing knowledge, such as describing, explaining, or comparing. Dalton-Puffer (2013) refers to Widdowson (1983) and Trimble (1985), who examined how a range of illocutions were realized and organized in academic and subject-specific texts, but also to lesser known approaches to communicative intentions in technical texts published in East Germany, namely Hoffmann (1988) or Gläser (1990), who work with the notion of *Kommunikationsverfahren* (Schmidt, 1981). However, Dalton-Puffer (2013) reports that their analyses do not provide concrete specifications and do not discuss pedagogical implications and subsequently points towards the notion of *functional pragmatics*, developed by Ehlich and Rehbein (1986). Looking into how language is used in institutions like schools, Ehlich and Rehbein (1986) observe that communication at school is a “highly dense and rarely interrupted sequence of talk” (p. 1, translated by Dalton-Puffer, 2013, p. 231) and that spoken interaction determines school like nothing else. As a consequence, communication at school promotes the formation and ritualization of purposeful verbal action patterns (*sprachliche Handlungsmuster*), such as justifications or narrations (Ehlich & Rehbein, 1986).

Based on the theories outlined above, Dalton-Puffer (2013) conceptualized CDFs as language patterns “which have crystallized in response to recurrent situative demands in a context where participants have recurrent purposes for communicating” (p. 231). As such, her criteria for structuring and condensing the discourse functions found in the literature are underlying communicative intentions concerned with the expression and sharing of cognitive processes, resulting in a construct consisting of seven types, namely CLASSIFY/ CATEGORIZE<sup>14</sup>, DEFINE, DESCRIBE, EVALUATE, EXPLAIN, EXPLORE, and REPORT.

---

<sup>14</sup> CLASSIFY was later renamed into CATEGORIZE as this label seemed more relevant for various subjects, including history (see also Dalton-Puffer & Bauer-Marschallinger, 2019). Accordingly, this thesis will use CATEGORIZE from now on (see also subsection 3.4.4.1).

### 3.4.3 Features of the CDF construct (Dalton-Puffer, 2013)

Table 3 below presents the CDF construct, illustrating that each CDF type rests upon a communicative intention relevant for the cognitive engagement with content knowledge (left column) and can be prompted via various CDFs verbs (right column). These CDF verbs present examples of action-related verbs and learning objectives often found in content subject curricula and materials. As can be seen in Table 3, the number of CDF verbs are not equal, but this list is not meant to be exhaustive. Moreover, it also depends on the discipline which CDF verbs might be more central or how they are understood exactly. Here, it should be noted that while the underlying communicative intention designates the core meaning, the specific realization and shape of the different types and their members depend on the discipline, making the categories somewhat fuzzy.

Table 3. The CDF construct (Dalton-Puffer & Bauer-Marschallinger, 2019, p. 35)

communicative intention	type	examples of CDF verbs
I tell you how we can cut up the world according to certain ideas	CATEGORIZE	<i>classify, compare, contrast, match, structure, categorize, subsume</i>
I tell you about the extension of this object of specialist knowledge	DEFINE	<i>define, identify, characterize</i>
I tell you details of what I can see (also metaphorically)	DESCRIBE	<i>describe, label, name, specify</i>
I tell you what my position is vis a vis X	EVALUATE	<i>evaluate, judge, argue, justify, take a stance, critique, comment, reflect</i>
I tell you about the causes or motives of X	EXPLAIN	<i>explain, reason, express cause/effect, deduce, draw conclusions</i>
I tell you something that is potential (i.e., non-factual)	EXPLORE	<i>explore, hypothesize, predict, speculate, guess, estimate, simulate</i>
I tell you sth. external to our immediate context on which I have a legitimate knowledge claim	REPORT	<i>report, inform, recount, narrate, present, summarize, relate</i>

A certain degree of blurriness also applies to the boundaries between the prototypical types, as these are neither completely disjointed nor are they mutually exclusive. For example, a complete definition would always include a classification and some classifications could entail a definition of a member. In fact, very often, CDFs build on and complement each other, creating tight nets of CDFs (see, e.g., Breeze & Dafouz, 2017; Dalton-Puffer et al., 2018; Dalton-Puffer & Bauer-Marschallinger, 2019). Lorenzo (2017) describes this as *functional stress* and argues that accumulating several functions into one proposition is a sign of sophistication. As a reaction, Dalton-Puffer et al. (2018) argue that CDFs run on two levels, namely on the level of *episodes*, being larger stretches of speech serving one overall communicative intention, and *basic* elements, constituting smaller CDFs sustaining and supplementing the overall purpose of the episode. Unlike Kidd (1996), who similarly grouped academic language functions into either macro- or microfunctions, Dalton-Puffer et al. (2018) do not determine which CDFs can function on the



lower or the upper level. In general, the types are not hierarchically structured according to complexity, considering that the level of complexity is dependent on the context of the operation and that there is no scientific basis that would determine such a sequence. Moreover, it should be mentioned that the labels used for the prototypes are common English verbs. Being aware that such labels create unstable and also flexible meanings, Dalton-Puffer (2013) explains that newly coined labels would not be appropriate as this would imply a certainty and clear definition of the underlying cognitive processes that are not warranted at this point. Finally, a strong presence of verbs (types and examples) underscores the actional nature of CDFs, reflecting the idea of speech acts.

In summary, the construct is flexible, non-essentialist to an extent that could be described as fuzzy. Yet, Dalton-Puffer (2013) argues that this ‘fuzziness’ is by design in order to permit the accommodation of various cultural and subject-specific educational requirements and models. By allowing the adjusting of some parameters of the construct from various disciplinary lenses, the construct might work as a heuristic in diverse contexts, enabling a common notion for discussion.

### 3.4.4 The 7 CDF types

#### 3.4.4.1 CATEGORIZE/ CLASSIFY

This CDF type is usually regarded as a central element of the academic discourse of various disciplines (e.g., in Beacco, 2010; Kidd, 1996; Mohan, 1986, but see Dalton-Puffer, 2013), as it relies on structuring and organising content, i.e., “how we can cut up the world according to certain ideas” (Dalton-Puffer, 2013, p. 234). In L. W. Anderson and Krathwohl’s (2001) taxonomy, categorizing/classifying belongs to *understand* since it relies on the identification of relevant characteristics and recognizing patterns, requiring a good understanding of the subject matter. Concerning the structure, for Trimble (1985), who also underlined the importance of this discourse function for academic discourse, a complete classification includes “the item (or items) being classified”, “the class to which the items (members) belong” and “the basis (or bases) for classification” (p. 86). For this process, one usually needs to compare and contrast, looking for similarities and differences (see also Evnitskaya, 2019). According to Trimble (1985), classifications can also be *partial* or *implicit*. Partial classifications omit the basis of classification, i.e., just identifying class-membership or recognizing members, whereas in implicit classifications, classifying information is presented without the use of terms typical for classifying. Trimble (1985) observed that learners tend to struggle with the direction of the classification (category → members vs. members → category, see also Dalton-Puffer, 2016). In terms of linguistic realization, classifications often come in the form of “‘X is Y’, ‘X is a member of Y’ or ‘X forms part of class Y’ for classifications” when members are assigned to class and “‘Y comprises X and Z’ for top-down classifications” (Evnitskaya, 2019, p. 241).

In the subject history, complete classifications in the narrow sense do not seem to be as typical as in the natural sciences. While historical discourse might include establishing temporal or social categories or categorizing source types, operations such as comparing, contrasting, and matching

are much more common, e.g., when corroborating sources or comparing past and present developments or values (Bauer-Marschallinger, 2016; Lorenzo, 2017). On the basis of conceptual and empirical exploration, Evnitskaya and Dalton-Puffer (2020), too, argue that classifying is more relevant in the natural sciences, whereas comparing is more predominant in history education, and thus they suggest using the label CATEGORIZE as superordinate term for two co-hyponyms CLASSIFY and COMPARE. Moreover, from an epistemological point of view, categorizations are assumed to be more flexible and often bound to the context, trying to group members based on perceived similarity/difference, in contrast to classifications which imply more fixed boundaries and systematic classes (see Ellin, 2004). For these reasons, the label CATEGORIZE is used in this thesis, as it appears to be more appropriate for the subject history and also covers comparing while not excluding complete classifications.

#### **3.4.4.2 DEFINE**

DEFINE organizes and structures knowledge by clearly establishing “the extension of this object of specialist knowledge” (Dalton-Puffer, 2013, p. 234). Formal definitions can be summarized by the formula *Species (term) = Genus (class) + Differentia (differences)* (Trimble, 1985, pp. 75–76). This shows that DEFINE usually entails CATEGORIZE but focuses on the member rather than the class. Trimble (1985) adds that variations in shape or size are possible. For example, *class* can be omitted, resulting in so-called *semi-formal definitions*, or they could be realized via synonyms or antonyms, which Trimble (1985) labels as *non-formal definitions*, reducing overall complexity. Conversely, definitions can also be made more complex and richer by adding other types of information, like limitations, descriptions, examples, instruction, explication, or further classification (Trimble, 1985), potentially overlapping with a number of CDF types apart from CATEGORIZE.

Given its close relations to other CDF types, defining plays a central role in academic discourse, as highlighted in many other constructs of academic language functions (e.g., Beacco, 2010, or Vollmer, 2010, but see Dalton-Puffer, 2013). As for historical literacy, Nashaat-Sobhy and Llinares (2020,) argue that defining is central in history because it pursues “the goal of studying the past in a systematized fashion” and for that, defining terms and concepts functions as crucial building blocks and “important stepping-stone[s] to perform higher-order functions” (pp. 2-3). Maset (2015), a history educationalist, similarly argues that learners can only engage in historical thinking if they understand and can work with key historical concepts and terms and are aware of their extension.

#### **3.4.4.3 DESCRIBE**

Describing refers to the process of providing details of what one perceives. This can include any “observable features, qualities or external and also internal characteristics of something [...] that can be a given object, entity, person, situation, event, or process” (Dalton-Puffer, 2016, p. 38). Providing details and making features obvious to others often forms a crucial step in constructing

content knowledge, and therefore DESCRIBE is usually considered to be a central academic language function (e.g., Beacco, 2010; Vollmer, 2011, but see Dalton-Puffer, 2013).

In the context of science, Trimble (1985) differentiated between three different types, namely *physical*, *process*, or *functional* description. Physical descriptions cover size, shape, material, colour, texture, or any other outward features, whereas process descriptions deal with procedures, i.e., sequences of steps and their goal (Trimble, 1985). Functional descriptions refer to the purpose and mechanics of an object and its individual parts, which might include descriptions of cause and effect, somewhat overlapping with EXPLAIN. As for historical discourse, descriptions are required to contextualize historical events by describing historical persons, inventions, etc., or to provide details about historical sources or artefacts (Bauer-Marschallinger, 2016; Lorenzo, 2017).

#### 3.4.4.4 EVALUATE

Considering that virtually all modern learning cultures subscribing to an enlightened worldview aim at enabling learners to take a personal stance on the basis of careful consideration and critical thinking, it is not surprising that one finds this thinking skill and a plethora of related performative verbs in many curricula, including Austrian curricula (e.g., Austrian Federal Ministry for Education, 2004, 2014, 2015a). According to Dalton-Puffer (2016), verbs like “appraise, argue, assess, bring evidence, check, critique, content, corroborate, debate, defend, evaluate, judge, justify, take a stance, refute, raise objections” (p. 41) all share the common underlying communicative intention of taking a stance on the basis of evidence, knowledge, experiences, or values.

In view of its central role in education, EVALUATE has been considered in many frameworks of academic language functions or thinking skills (e.g., L. W. Anderson & Krathwohl, 2001; Beacco, 2010; Bloom et al., 1956; Mohan, 1986; Vollmer & Thürmann, 2010, but see Dalton-Puffer, 2013). In the taxonomy of thinking skills (L. W. Anderson & Krathwohl, 2001; Bloom et al., 1956), EVALUATE is regarded as a higher-order thinking skill, which involves “making judgments based on criteria and standards” (L. W. Anderson & Krathwohl, 2001, p. 83). These criteria or standards can be quantitative, i.e., expressible in numbers, and qualitative, i.e., non-numerical evidence, criteria, standards, or moral judgements (L. W. Anderson & Krathwohl, 2001; Dalton-Puffer, 2013).

As for historical literacy, Lorenzo (2017) argues that “[e]valuation represents the endpoint in historical discourse maturation” (p. 37), which entails a number of relevant historical skills, including *corroboration*, *stance-taking*, and *heteroglossic critique*:

- Corroboration refers to the ability to contrast views or sources against facts or other sources (see also Maset, 2015).
- Stance-taking involves judging the past without stereotyping or dualistic thinking. Furthermore, it involves assessing the impact of historical developments or values on the present and one’s own historical identity or awareness (see also *Orientierungskompetenz* in Körber et al., 2007, and section 4.2.3).

- Heteroglossic critique describes the process of forming a personal stance on the basis of a variety of sources (see also *Rekonstruktionskompetenz* and *Re-Organisationskompetenz* in Körber et al., 2007, and section 4.2.2 and 4.2.3).

According to Lorenzo (2017), these history skills, embodying the CDF type EVALUATE, are central to historical discourse, signifying “the full development of a voice” (p. 37; see also Coffin, 2006). From a history-didactics perspective, too, many competency models stress the importance of evidence- and criterion-based evaluation of sources, views, or past developments as well as the significance of taking a stance in relation to one’s own historical identity or perceptions of the world (see Bauer-Marschallinger, 2016, and chapter 4).

#### **3.4.4.5 EXPLAIN**

EXPLAIN is a frequently used verb in everyday communication but also in educational contexts and thus expresses a variety of related but different meanings. Referring to the *Oxford English Dictionary*, Dalton-Puffer (2016, p. 44) provides three different meanings of “explain”, namely:

- *Explain 1: To make sth. plain or intelligible; to clear of obscurity or difficulty; to give details of or to unfold (a matter)*
- *Explain 2: To give an account of one's intentions or motives*
- *Explain 3: To make clear the cause, origin, or reason of*

As the first understanding of EXPLAIN is very general and broad, rather overlapping with detailed description, Dalton-Puffer (2013, 2016) excludes this meaning. Instead, this CDF type is based on meaning 2 and 3, which both share a focus on causality, once focusing on human intentions, central in humanities and social sciences, and once centring on deductive explanations of phenomena or developments, more relevant for natural sciences. As such, the underlying communicative intention of this type is defined as “I give you reasons for and tell you about the cause/s of X” (Dalton-Puffer, 2013, p. 234). This understanding is also consistent with the conceptualization of EXPLAIN in the taxonomy of thinking skills (L. W. Anderson & Krathwohl, 2001; Bloom et al., 1956), where explaining is part of *understand* and defined as “construct[ing] and us[ing] a cause-and-effect model of a system” (L. W. Anderson & Krathwohl, 2001, p. 75).

Considering that the question ‘why?’ usually plays an important role in constructing disciplinary content, it is not surprising that many previous frameworks include this CDF type (e.g., Biggs & Tang, 2011; Kidd, 1996; Vollmer & Thürmann, 2010, but see Dalton-Puffer, 2013). As for history education, as already shown in subsection 3.2.2.2, cause-effect relations shape and define historical discourse. Ideally, as learners mature, simple chains of cause and effect evolve into multifactorial, complex explanations, realized via grammatical metaphor and asyndetic structures (Achugar & Schleppegrell, 2005; Lorenzo, 2017).

#### **3.4.4.6 EXPLORE**

This CDF type refers to any verbal actions in which we share something “that is potential. i.e., non-factual” (Dalton-Puffer, 2013, p. 234), meaning anything “not in the here and now, and which is not firmly established past fact either” (Dalton-Puffer, 2016, p. 46). Examples of such operations

include assuming, predicting, speculating, conjecturing, supposing, or hypothesizing, which all occur in each other's lists of (near-) synonyms (Dalton-Puffer, 2016). This relatively broad list of CDF verbs also shows that the meaning of EXPLORE is not limited to notions of hypothesis and prediction in a strictly scientific sense but rather denotes a non-technical, semi-expert understanding of these processes. Concurring with many other frameworks of academic language or thinking skills, such as Biggs and Tang (2011), Beacco (2010), Kidd (1996), or Vollmer (2010), Dalton-Puffer (2016) argues that exploring plays an essential role in constructing disciplinary knowledge, which is typically regarded as a complex, higher-order thinking skill. L. W. Anderson and Krathwohl (2001), for instance, assign generating hypotheses or alternatives to the highest category, namely *create*. Linguistically, EXPLORE often relies on complex lexicogrammatical structures, including modal verbs, adverbs, and conditional clauses (Dalton-Puffer, 2016).

In historical discourse, EXPLORE seems to play a minor role at first glance as future predictions are rare and thought experiments are often not considered to be appropriate in historical discourse (Lorenzo, 2017). Quoting Hobsbawm (1997, p. 150), Lorenzo (2017) maintains, however, that “all history is full of implicit or explicit counterfactuals, ranging from speculations about alternative outcomes to more specific might-have-beens” (p. 38), so students of history still need to be able to speculate about alternative, imaginative scenarios and express counter-factuality. Moreover, there are many aspects historians cannot fully determine, especially if sources are incomplete, potentially unreliable, or if one simply does not have access to a full range of historical evidence as would be the case in typical school history lessons (Bauer-Marschallinger, 2016, see also *Dekonstruktionskompetenz* in Körber et al., 2007 or section 4.2.2). In these cases, it is vital to express that other interpretations are possible. On a more abstract level, learners are also required to EXPLORE their own historical identity and other people's perspectives, which are intangible constructs one can only delve into via conjecture (Bauer-Marschallinger, 2016, see also *Orientierungskompetenz* in Körber et al., 2007 or section 4.2.3).

#### **3.4.4.7 REPORT**

This CDF type covers all instances when one informs somebody of “what happened, when, who did it and to whom under what circumstance” (Dalton-Puffer, 2016, p. 49), summarized as “I tell you sth. external to our immediate context on which I have a legitimate knowledge claim” (Dalton-Puffer, 2013, p. 234). Examples of CDF verbs denoting this function include “inform, recount, narrate, present, summarize, [or] relate” (Dalton-Puffer, 2013, p. 235). These operations all “assume a reduced shared background knowledge of speaker and recipient” and involve careful selection of information to pass on to the audience (Dalton-Puffer, 2016, p. 49). Academic language functions or thinking skills encapsulating these meanings can be found in many frameworks (e.g., Biggs & Tang, 2011; Kidd, 1996; Vollmer & Thürmann, 2010, but see Dalton-Puffer, 2013).

In school, this often happens after learners work on a topic and then report back to class, e.g., in the form of an oral presentation, or to the teacher, which typically happens in the written form, which, depending on the subject, might constitute established genres in their own right, such as a

lab report in science or a historical recount in history (Dalton-Puffer, 2016). In these cases, one might also report and summarize other CDF types, making the CDF type REPORT prone to functioning as episode-CDF.

Zooming in on historical discourse, according to Lorenzo (2017), “[r]eporting is the historical function par excellence and recalls the original meaning of history as the past witnessed”, with its textual form of the narrative (p. 38; see also *recounts* in Coffin, 2006). This is reflected in many models of historical competences, which consider *narrative competence* as key element (e.g., Gautschi, 2015; Pandel, 2012; Rösen, 1983). These historical narratives are understood as reconstructions of the past on the basis of careful investigation and deconstruction of historical sources while also considering the historical context, which is why other models use the term *reconstruction competence* (see Körber et al., 2007, or section 4.2.2). Some also argue that the difference between narrating and reporting lies in their degree of objectivity, with narrations implying more subjectivity than reporting (Dalton-Puffer, 2016; Vollmer, 2011). In historical narration, any recount involves the reconstruction of the past, which ideally is based on careful consideration and analysis of sources. As such, a historical narrative always involves subjective judgement (e.g., which elements to include, which ones to leave out, and how to link *participles of the past*, see section 4.2.2), and can thus only aim for intersubjective comprehensibility (Kühberger, 2015). For these reasons, historians tend to consider this operation as complex (e.g., Kühberger, 2011) unlike more general frameworks, such as L. W. Anderson and Krathwohl (2001), who put it into *understand*, i.e., one of the lower-order thinking skills.

### 3.4.5 Empirical validation of the construct

Since the CDF construct is based on literature which, for the most part, rests on educational standards and curricula, Dalton-Puffer (2013) argues that further empirical validation needs to take place. One main area of investigation would be observational-descriptive studies in a variety of subjects examining if and how the construct reflects the reality of subject-specific knowledge construction, exploring how and by whom CDFs are realized or whether there is a meta-level. Moreover, Dalton-Puffer (2013) concurs with Bailey and Butler (2003) that a construct that only considers educational standards and curricula runs the risk of being somewhat detached from the reality of school, representing more of an idealist version of the language to be used in school, and consequently teachers might not find it useful. Therefore, Dalton-Puffer (2013) calls for empirical research that investigates the usefulness and face validity of the construct from the teachers’ perspective.

Most of the empirical studies conducted since then have been of the observational type. Summarizing five MA theses supervised by the author of the CDF construct, Dalton-Puffer et al. (2018) could show that all seven CDF types were featured regularly in classroom interaction across five different subjects, namely physics, biology, business/economics, history, and EFL. The most frequently produced CDF type was DESCRIBE, being the most common CDF type in all subjects except business/economics, where REPORT was the most frequent type. EXPLAIN and DEFINE were

also very common in the data of the five studies. The distribution of the other CDF types, however, seemed to depend on the discipline. For example, while instances of EVALUATE were relatively prominent in social sciences and humanities classes, they were rather rare in natural science lessons. In turn, EXPLORE was much more common in the biology and physics lessons than in business/ economics, history, or English.

Looking into who performed these CDFs, Dalton-Puffer et al. (2018) could show that in the natural sciences, students rarely performed CDFs on their own. Instead, these were either produced by the teachers alone or co-constructed with learners. Conversely, in business/economics and especially in English lessons, learners realized considerably more CDFs autonomously, indicating higher degrees of learner-centredness and more opportunities for learners to verbalize their cognitive processes. Three of the five studies also investigated the occurrence of meta-talk, i.e., “talk about talk” (Lemke, 1990, p. 118), reporting that CDF-related notions are sometimes, but not frequently, explicitly addressed both reactively and proactively, usually by talking about performative verbs or nominal derivatives. However, the classroom data reported in Dalton-Puffer et al. (2018) showed that performative verbs were not always used in a precise way. For example, EXPLAIN tended to be used as dummy CDF, covering anything where a detailed description was called for, reflecting the non-technical explain-1 meaning mentioned on page 62. Additionally, Dalton-Puffer et al. (2018) provide some insights into the lexicogrammatical realization of CDFs. The main tenor across the data sets seems to be that learners rarely constructed complete CDFs autonomously and if they did, they only used basic markers, whereas explicit markers that would signify their communicative intention were rare exceptions. For example, students added “maybe” or “I think” to their utterance to signal hypotheticality (EXPLORE) rather than using conditional auxiliaries. Finally, on the conceptual level, the analysis of the five data sets has shown that CDFs operate on two hierarchical level, i.e., basic CDFs and episode CDFs (see section 3.4.3).

Based on Bauer-Marschallinger’s (2016) MA thesis focused on history education featured in Dalton-Puffer et al. (2018), Dalton-Puffer and Bauer-Marschallinger (2019) could show that the CDF construct is not only compatible with the competency model underlying the Austrian curriculum, but also that the seven types play a crucial role in classroom discourse and written productions by Austrian learners of history. Observing four lessons in lower and upper secondary history education each, Dalton-Puffer and Bauer-Marschallinger (2019) report that these students produced all CDF types while working on their historical competences.<sup>15</sup> In both contexts, DESCRIBE and EXPLAIN were particularly central for working with sources (*methodological competence*), whereas EVALUATE predominated when learners reflected and re-organised their historical awareness (*orientation competence*). Moreover, the datasets show a close relationship between factual competence<sup>16</sup> and DEFINE. Comparing the classroom data of the two age groups,

---

<sup>15</sup> EXPLORE was only realized in the upper secondary data set. All other CDF types appeared in both.

<sup>16</sup> In previous publications, I used the term *historical expertise*, but I have since realized that *factual competence* might be a less ambiguous translation.

Dalton-Puffer and Bauer-Marschallinger (2019) noticed that older students produced CDFs more frequently and could employ a greater variety of CDFs for the same historical competences.

After the four lessons observed, the students completed a competency-based written assignment in which they had to engage with historical sources and thereby, ideally, demonstrated all four competences featured in the curriculum. The individual items were formulated with performative verbs, targeting specific CDF types. In many cases, the students indeed realized appropriate CDF types. Interestingly, the learners could also, by and large, produce CDFs that were rarely observed in classroom interaction, such as EXPLORE or CATEGORIZE. However, comparing younger and older learners, Dalton-Puffer and Bauer-Marschallinger (2019) observed that younger students struggled with employing appropriate CDF types more frequently. For example, EXPLAIN was often substituted with seemingly easier CDF types such as DESCRIBE. However, it is not clear whether these learners interpreted the prompt incorrectly or whether they were linguistically or cognitively unable to express the target function. Unsurprisingly, the older students also realized CDFs more proficiently by providing better logical-semantic connections and presenting more structural and lexical variety than the younger learners. However, both classroom as well as written data showed that lower and upper secondary learners often did not signal their communicative intentions clearly, and when they did, they used a limited selection of basic markers. Based on these results, Dalton-Puffer and Bauer-Marschallinger (2019) recommend explicit instruction and some practice regarding the use of CDFs to help CLIL learners demonstrate their subject-specific skills more competently. Moreover, Dalton-Puffer and Bauer-Marschallinger (2019) conclude that “investigating tests and students’ answers through a ‘CDF-lens’ helps to grasp the core demand of an individual test item and comprehend the students’ process of reasoning” (p. 54), asserting the usefulness of the construct for test design and assessment but also for empirical research.

In Germany, Vanderbeke and Wilden (2017) analysed oral CDF use during lab sessions in the context of bilingual biology at upper secondary level. Here, the most common CDFs were EVALUATE and DESCRIBE, followed by EXPLAIN, EXPLORE, and REPORT, and, lastly, DEFINE/CATEGORIZE (merged in this study). Additionally, Vanderbeke and Wilden (2017) observed that learners often made use of the physical context, their peers, their materials, or the dictionary to cope with gaps in vocabulary but also to construct CDFs, for which the authors used the label *affordances*. They further noted that while learners used CDFs in peer-interaction, their CDFs were usually very short and simple. Thus, Vanderbeke and Wilden (2017) recommend using and highlighting CDF-related linguistic support, helping learners construct longer and more complex CDFs.

In Spain, Lorenzo (2017) analysed written historical narratives of 21 10<sup>th</sup>-grade CLIL students with a focus on which CDFs these learners produced and how they realized these linguistically. In this corpus, all CDFs were present, and some students managed to perform these CDFs in a way indicative of high levels of historical literacy. Based on the aspects of historical sophistication found in the data and established in the literature, Lorenzo (2017) defined *cognitive discourse competencies for advanced historical thinking* for all seven CDF types. These competencies have



been considered in the characterization of the seven types in section 3.4.4. For example, the learners participating in Lorenzo's (2017) study demonstrated the capability of producing multifactorial explanations and not just personalised, simple relations of cause and effect (EXPLAIN). As for REPORT, Lorenzo (2017) observed that some students were able to organize and contextualize information on a historical event or development, i.e., linking their narrations of one event to what they already knew about the historical context. In terms of categorizations, the corpus contained examples of *functional stress*, i.e., the merging of two or more functions, which can be interpreted as another sign of advanced historical literacy according to Lorenzo (2017). Linguistically, this often requires complex sentence structures and nominalisations. Concerning DESCRIBE and DEFINE, Lorenzo (2017) observed some examples where the communicative intention was explicitly marked, i.e., by stating "can be defined as" for DEFINE or "represent" or "characterized" for DESCRIBE (p. 37). He further reported that some students were able to include some abstractions (e.g., "futurism", "expressionism", "loss of sense", p. 37). However, Lorenzo (2017) added that these signs of advanced historical competence did not happen across the board. Moreover, he noticed that EXPLORE and EVALUATE were hardly found in the data. In the case of EXPLORE, this matches the results by Dalton-Puffer and Bauer-Marschallinger (2019) and Dalton-Puffer (2007), which might stem from the fact that hypothesizing is not as established as other functions in historical discourse (see 3.4.4.7). As for EVALUATE, Lorenzo (2017) observed that few learners took personal stances to begin with and those that did were often overtly biased and/or failed to provide sound justifications, "pass[ing] off opinion as fact" (p. 38). On the whole, however, Lorenzo (2017) argues that his results suggest that CLIL indeed "facilitates the acquisition of advanced historical functions and knowledge as represented by CDFs" (p. 40). Methodologically, he concludes that "[f]or those seeking a comprehensive but also manageable taxonomy to conduct research on language across the disciplines, Dalton-Puffer's classification is an accurate lens through which to examine classroom discourse and literacy studies" (p. 40), validating the construct for research purposes.

Also in Spain, the UAM-CLIL research group has been investigating the development of CLIL learners' academic language competence in low- and high-intensity tracks as these learners transition from primary to secondary education. Amongst other tools grounded in SFL and legitimation code theory (Maton, 2013), they used Dalton-Puffer's (2013) CDF construct to analyse spoken and written productions in the learners' L1 and L2 at various grade levels and subjects. In the context of history education, Nashaat-Sobhy and Llinares (2020) focused on written L1 and L2 definitions in two sub-fields of history (social groups and historical periods) by the same learners at grade 6 and 8. They found that at secondary level, when they had to define historical periods, the learners produced formal definitions more frequently than at primary level, when the focus was on social groups. Especially at primary level, students often omitted class terms in their definition, which might have to do with the primary textbooks used because these often leave empty the slot for class, particularly when defining periods (Nashaat-Sobhy & Llinares, 2020). Looking at the differentia and expansion of the learners' definitions, Nashaat-Sobhy and Llinares (2020) report that the secondary-level students' definitions presented more variety and

did not differ substantially between L1 and L2 productions. However, this corpus shows that primary learners opted more often to define in Spanish instead of English than at secondary level, suggesting that older learners might have gained more confidence. Yet, Nashaat-Sobhy and Llinares (2020) argue that the different fields (i.e., social groups and historical periods) and the way these were defined in the textbooks might have affected the way these learners defined. Consequently, Nashaat-Sobhy and Llinares (2020) argue that “it is necessary to operationalize definitions and identify types and components and how these vary across fields” (p. 11), especially since definitions are often used to assess the learners’ subject knowledge.

Linking the CDF type EVALUATE to the appraisal framework, Whittaker and McCabe (2020) examined students’ written and spoken evaluations in art, ecology, and history education at two points in time. Across disciplines, they observed that the appropriateness of field and evaluation couplings improved somewhat and recommend more explicit instruction concerning cognitive discourse competence via the appraisal framework. Zooming in on their results for the subject history at primary level, Whittaker and McCabe (2020) report that although learners are limited in their selection of evaluative language, e.g., by copying words from the prompt or using very basic vocabulary, most of them were able to justify their evaluations of historical events (*appreciation* in SFL terms) usually by providing basic information on the historical context or historical impact. At secondary level, some learners were able to provide rather objective and disciplinary evaluations of social groups (*appreciation*) coupled with justifications grounded in the historical context. However, some learners took the prompt too literally and expressed affect, i.e., personal valuations and emotions, which resulted in less appropriate answers.

Also working with the UAM-CLIL corpus, Evnitskaya and Dalton-Puffer (2021) examined oral categorizations in history and science education at grade 6. They found that comparisons were much more frequent than classifications across the corpus. In fact, classifications were almost absent from the history data. Moreover, the classifications produced were often incomplete but still formed a central element when talking about zoology or botany. Zooming in on comparisons, Evnitskaya and Dalton-Puffer (2021) observed that learners compared both on the basis of similarity and difference but seemed to favour difference-based comparisons, which were often implicit by juxtaposing two contrasting elements. As for the learners’ lexicogrammatical realizations of categorizations, Evnitskaya and Dalton-Puffer (2021) found that the learners’ lexical choices in the L2 were comparable to their L1 performance, if somewhat less varied. The authors further add that their results do not suggest a discrepancy between the students’ linguistic means in their L2 and their cognitive level.

Switching perspective to the teachers’ use of CDFs, Evnitskaya (2019) investigated classifications in classroom interaction from a multimodal perspective. The teacher participating in her case study made use of a variety of semiotic tools to support the students’ comprehension. To be more precise, this teacher combined paralinguistic and linguistic strategies, e.g., gestures, material objects, or visualisations, but also prosodic elements and switching languages or registers, in order to scaffold the learning process. As such, teachers and learners co-constructed

classifications, with the teacher's role of transforming the learners' "everyday wordings into appropriate school-science statements" (Evnitskaya, 2019, pp. 257–258). However, Evnitskaya (2019) noted that these processes of unpacking and repacking were done without explicit meta-talk concerning the CDF or academic language use in general, missing the opportunity to raise the students' awareness and provide learners with the appropriate lexicogrammatical resources to construct their own classifications.

Not part of the UAM-CLIL research group but also in the Spanish context, Breeze and Gerns (2019) investigated the impact of a writing module focused on general academic writing on secondary history learners' realizations of DESCRIBE and EXPLAIN. They found that before the intervention, about half of the learners realized their CDFs without any discourse markers signalling their communicative intentions and without academic paragraph structure. After the intervention, the learners used linking devices more frequently and, for the most part, more appropriately, resulting in a greater variety of linking devices. Breeze and Gerns (2019) argue that although "these improvements could perhaps be dismissed as superficial, [...] the rise in explicit signposting accompanies better overall organisation and greater attention to the communication of meaning" (p. 30). Interestingly, from a content perspective, the authors also noticed that the learners' productions were more complete, suggesting that the focus on language indeed helped them demonstrate their knowledge. Since the writing instruction did not focus on CDFs explicitly, Breeze and Gerns (2019) hypothesize that explicit metalinguistic instruction concerning CDFs might help learners even more, also in view of exam prompts, which often contain performative verbs.

A study looking into students' productions elicited by exam items containing performative verbs was conducted by Breeze and Dafouz (2017) in the tertiary context. In this study, they examined Spanish and English exam responses where students had to analyse an advertising campaign, linking visual evidence (DESCRIBE) and theoretical concepts in order to characterize the target group (CATEGORIZE) and EXPLAIN the role of emotions, attitudes, and motivations in this advertising campaign. Breeze and Dafouz (2017) found that low-level answers in both languages often failed to establish the connections between the visual elements of the source and the theoretical concepts. Moreover, they were often vague and lacked technical terminology. As the teacher of this class put it, "[i]n many cases their sentences can't really be understood. You can grasp intuitively what they might have meant if they had expressed themselves clearly" (Breeze & Dafouz, 2017, p. 88). In contrast, high-level answers signalled purpose and appropriately linked concrete and relevant features of the source material to theoretical concepts (DESCRIBE-CATEGORIZE) or to effects on the target audience (DESCRIBE-EXPLAIN). Comparing L1 and EMI performances, Breeze and Dafouz (2017) found no considerable differences, suggesting that at least for some, the issue might be misconceptions of what certain questions entail and require, both in terms of content and linguistic realization. They therefore suggest that lecturers should unpack difficult task requirements and provide models and examples of how certain functions can be enacted and, more importantly at this level, how different CDFs can be connected.

Also in the tertiary context, Doiz and Lasagabaster (2021) examined history lectures. They found that the three lecturers involved used complex CDFs, combining different smaller CDFs to sustain larger communicative intentions, similar to what Breeze and Dafouz (2017) observed in high-level student performances. Zooming in on different central CDF types, Doiz and Lasagabaster (2021) report that the least-proficient lecturer expressed causality mostly via prototypical cohesive devices such as “that’s why” or “because” (p. 63). More proficient lecturers, on the other hand, expressed causality within a clause rather than between clauses, e.g., by using phrases such “led to” or “this paves the way” (Doiz & Lasagabaster, 2021 p. 63) as has been described by, e.g., Achugar and Schleppegrell (2005) or Schleppegrell et al. (2004) (see subsection 3.2.2.2). As for DEFINE, Doiz and Lasagabaster (2021) found that the lecturers liked to use an inclusive “we”. Overall, these lecturers used all CDFs but not to an equal extent, with DESCRIBE, REPORT, and EXPLAIN being the most common ones and CATEGORIZE and EXPLORE the least present ones. Conceptually, Doiz and Lasagabaster (2021) argue that EVALUATE should explicitly include reported evaluations and that DEFINE, DESCRIBE, and REPORT need to explicitly include a temporal aspect as well. These points will be revisited in the conceptual discussion in subchapter 8.2. Pointing towards future research, Doiz and Lasagabaster (2021) acknowledge that the analysis was done by two applied linguists and thus call for co-validation of such analyses by history experts to truly integrate the two perspectives.

Focusing on content-and-language-integrative, classroom-based assessment, deBoer (2020) examined how CLIL students at tertiary level use CDFs when interacting and explored how these interactions might be used for determining the learners’ abilities. He found that even low-proficiency learners (approximately A2) were able to interact and employ a range of CDFs when working on content in a meaningful way. According to deBoer (2020), these results imply that teachers can utilize the notion of CDFs to dynamically assess the students’ understanding, for instance, by asking learners to elaborate on their descriptions, explanations, etc.

### **3.4.6 Operationalizing the CDF construct**

Many of the studies outlined above conclude with a call for operationalizing the CDF construct to help learners appropriately realize CDFs and improve their academic and subject-specific literacy skills. According to Meyer et al. (2015), the students’ inability to use CDFs is why CLIL “has yet to live up to its full potential” (p. 44), but so far, little has been done to support learners in this endeavour. Morton (2020) agrees and maintains that CDFs can function as “building blocks” to ensure “focused and principled integration of content, literacy, and language” (p. 11). He further argues that these building blocks allow clearly formulated learning objectives that are pertinent to the target content and language but can also be used for providing literacy-related formative feedback or summative assessment in a tangible and relevant way. Thus, Morton (2020) puts forward that now one needs “to ‘translate’ what we know about CDFs into effective and practical instructional and assessment strategies” (p. 16). Thus far, the number of completed studies aiming for such translations into successful classroom practice is limited, but currently, a handful of PhD studies designing classroom materials on the basis of CDFs are underway.

Starting with completed studies, Nashaat Sobhy (2018) operationalized defining for an undergraduate course she taught at a Spanish university. In this action research project, Nashaat Sobhy (2018) created activities which should help her own students move gradually from mundane, context-embedded, and cognitively undemanding to subject-specific, context-reduced, and cognitively demanding definitions (see Cummins' (1984) quadrant model for adjusting language) via guided attention, scaffolded in- and out-of-class practice, and the provision of a language support file. Nashaat Sobhy (2018) reports that most students experienced defining as challenging yet important and thus welcomed the intervention. The students found the intervention useful and felt that they could also retain content knowledge more effectively. Looking at the students' performance, Nashaat Sobhy (2018) observed a spectrum of successful and less successful definitions. One linguistic issue connected to defining (and academic discourse in general) that stood out was nominalisation, as many of her students tended to avoid nominalised structures and abstract language even after the intervention, making their definitions appear less academically appropriate and credible. Therefore, she recommends paying more attention to nominalisations and abstraction in future materials. From her perspective as instructor, Nashaat Sobhy (2018) found Dalton-Puffer's CDF construct useful as a "'blueprint' to teaching defining" (p. 108), as it clarified linguistic and content-related expectations. She thus recommends "that these CDFs would be explored in more depth and breadth by both content and language teachers alike, and in different disciplinary genres, as means to conciliate teaching content and language" (Nashaat Sobhy, 2018, p. 110).

Focusing on EXPLAIN, Connolly (2019) conducted an intervention study in lower and upper secondary chemistry education. For this intervention, Connolly (2019) created materials that scaffold cause-effect relations, following the pluriliteracies approach (see subchapter 0). These materials provide explicit instruction and opportunities to practise explaining in a guided way. Adopting a quasi-experimental methodology, Connolly (2019) focused on the effect of this intervention rather than on the features of CDF-based materials in chemistry. She found that the experimental group improved more than the control group both on the content knowledge and subject literacy scale as well as in terms of self-confidence. Affective factors remained unchanged for both treatment and control group.

Being interested in helping teachers effectively plan and conduct scaffolded content-and-language-integrative lessons, Tedick and Lyster (2019) offer practical tools that consider and expand the CDF construct. For example, the authors suggest a six-step procedure for writing language learning objectives, starting with (1) determining the discourse type, then (2) identifying the communicative functions needed for the activity in question, followed by (3) clarifying what language is expected for said function, including (4) grammatical structures, and (5) vocabulary, and finally, (6) using these insights to formulate language objectives. For this second step (communicative function), Tedick and Lyster (2019) refer to Dalton-Puffer's (2013) CDF construct but add an eighth type called INQUIRE, which includes the sample verbs "ask, examine, request, solicit, query, question, quiz" (Tedick & Lyster, 2019, p. 214). For them,

academic language functions are “content-obligatory language” (p. 93), and as such, these need to be reflected in learning aims and communicated to the learners. Coming back to their CAPA model (*contextualization, awareness, practice, and autonomy*, see p. 32 of this thesis), Tedick and Lyster (2019) argue that academic language functions are a crucial element in counterbalanced instruction and are especially useful during the *autonomy* phase of the CAPA sequence.

Turning to ongoing studies, one example focusing on history education is provided by del Pozo (in progress), who conducted an intervention study for which she created a CDF-based history module aimed at enhancing historical critical thinking skills of upper secondary bilingual students (see also del Pozo, 2019). This semester-long project followed a competency-based approach which should enable learners to deconstruct historical sources and, subsequently, build up historical narratives. According to del Pozo (2019), various inquiry tasks created in this project focus on different CDFs and, as such, her planning tools include CDF-based learning goals criteria. In her PhD study, del Pozo (in progress) does not focus on how the concept of CDFs is used in the classroom but on assessment of student performances, developing an analytical framework for integrated assessment (see also del Pozo & Llinares, 2021). First results suggest that the CDF construct is a useful tool to conceptualize the integration of content and language learning for assessment purposes since her rubrics provide viable ways for the assessment *of* and *for* learning (del Pozo & Llinares, 2021). Yet, the authors conclude that more research is needed concerning integrative approaches to assessment in CLIL.

Focusing on science education, Hasenberger (2018, in progress) is developing a CLIL module based on the concept of CDFs in the context of upper secondary science education. Having identified central disciplinary concepts and their connected CDFs, Hasenberger is developing teaching cycles that consist of (1) *pre-instruction activities* activating prior knowledge, (2) an *input-phase* providing direct instruction concerning content and the linguistic realization of relevant CDFs, and (3) *post-input* activities where learners are supposed to produce these CDFs in content-based, interactive tasks. Preliminary results of first research cycles suggest that the students perceive a learning benefit both in terms of science learning and language and that actual performance seems to improve as well, reflected in increased and more appropriate CDF realizations in the post-input phase (Hasenberger, 2018). Gerns (in progress), too, is focusing on CLIL for the natural sciences and has developed lower secondary teaching materials which provide explicit instruction on discipline-appropriate written realizations of comparisons.

To conclude, although many researchers who investigate CDFs theoretically and empirically have called for operationalizing this construct for classroom use, only few studies are available that actually do this. Yet, considering that the majority of the studies presented here are either recent or still ongoing, more are to be expected. As argued in chapter 2 and 3, such studies are indeed needed in the various disciplines and on different educational levels. The present PhD study intends to add to this growing research field by creating CDF-based competency-oriented history materials for upper secondary CLIL learners in collaboration with secondary teachers via multiple design and evaluation cycles.

## 4. History education

*"History is who we are and why we are the way we are."* (McCullough, 1984)

This quote by author and historian David McCulloch reflects current conceptualizations of what it means to learn about history. While in the past, history education focused on passing on historical knowledge and so-called *master narratives*, history education today aims at the development of historical consciousness and critical reflection on our own historical identity. The following chapter zooms in on current central notions and themes in history didactics (4.1) and, more specifically, how history learning is conceptualized in the broader context of this study by reviewing the competency model underlying Austrian secondary history curricula (4.2). In subchapter 4.3, the context of this study is explored further by providing an overview of the Austrian school system (4.3.1), the student body of higher vocational schools in general and in the school type featured in this study, i.e., the secondary college of business administration (4.3.2), and the specifics of history education in this school type (4.3.3).

### 4.1 Central notions and themes in history didactics

This subchapter discusses key concepts of history didactics that are currently debated in the research community and often reflected in today's history curricula, namely *historical competence* (4.1.1), *historical thinking* (4.1.2), *historical consciousness* (4.1.3), *historical literacy* (4.1.4), and *historical reasoning* (4.1.5). Van Drie and van Boxtel (2008) point out that some researchers in the field use some of these terms synonymously while others emphasize their difference. In any case, as the following sections will show, many of these notions are interdependent and all of them point towards an understanding of history education that highlights critical thinking and subject-specific skills rather than accumulating historical knowledge, calling for competency-based approaches to history learning.

#### 4.1.1 Historical competence

History education, especially in the German-speaking world, has recently undergone a paradigm shift following the so-called *PISA shock*, when large-scale standardized testing (PISA) revealed the shortcomings of knowledge-based curricula (Gautschi, 2015; Kölbl & Konrad, 2015; Kühberger, 2015). Traditionally, secondary history education was conceptualized as knowing facts and dates and recounting historical events and developments, reflected in curricula mostly listing topics (e.g., "Wiener Kongress – Vormärz - liberale Strömungen — Revolution von 1848" Austrian Federal Ministry for Education, 1995, p. 7637) and only containing receptive learning aims (e.g., by using various synonyms of "understand" and "gain insights", Austrian Federal Ministry for Education, 1995, own translation). As a consequence, traditional history education, at least in the German-speaking world, reproduced knowledge and certain master-narratives, which resulted in a very narrow understanding of history education, putting into question the purpose of this subject (Gautschi, 2015; Körber & Meyer-Hamme, 2015). Despite initial concerns that competency-based approaches, in general, could reduce education to outcome-oriented

mechanisms eventually serving the labour market, competency-based approaches were developed and integrated into national curricula of various subjects, including history, in the hope of improving educational practice and outcomes (Heil, 2012). In the context of history, this meant that history education was conceptualized beyond factual knowledge, fostering critical thinking skills as well as subject-specific and transferable competences (Gautschi, 2015; Körber & Meyer-Hamme, 2015; Kühberger, 2015).

The term *competence* is central in many disciplines and thus presents different underlying assumptions and understandings depending on the domain. Educational models, especially in the German-speaking context, are often based on Weinert's notion of competence:

In accordance with mainstream thought, we understand 'competence' as referring to combinations of those cognitive, motivational, moral, and social skills available to (or potentially learnable by) a person or a social group that underlie the successful mastery through appropriate understanding and actions of a range of demands, tasks, problems, and goals. (Weinert, 2001, p. 2433)

Weinert (2002) further explains that being competent also implies not only the ability to solve problems in theory but to be willing and ready to apply these skills in a responsible way in various contexts. Jung (2010) adds that these competences are not just relevant in the safe space of the classroom but also in real-life contexts, which often might be complex and disorganized, reflecting the etymology of the word "competence", which is "empowerment to cope" (p. 1). In contrast to this very general notion of competence, Pandel (2017) and Klieme and Leutner (2006) argue that competences are domain-specific, meaning that learners might be very competent in one field-specific area and less competent in another. This also implies that historical competences might not be dependent on general problem-solving skills (Pandel, 2017). For Pandel (2017), competences are domain-specific problem-solving skills, which can be described on different levels of quality and which are creative and generative in nature, containing multiple elements, such as skills, abilities, routines, techniques, and understanding of field-specific conventions. Kühberger (2015) also understands competences as problem-solving skills which enable learners to make use of their previous knowledge and set of skills and thereby allow the learners to improve their skills further and to adapt them for new domains. As such, this process can only come from the learners themselves, with (self-) reflection being the prime driving force (Gautschi, 2015). Structurally, Pandel (2017) explains, competences lie between the development of historical consciousness, i.e., the overarching goal of history education, and the concrete tasks used in school.

#### **4.1.2 Historical thinking**

The notion of *historical thinking* plays a central role in the practice of history teaching in a number of different learning and teaching traditions, including the German- and English-speaking world, and has thus been approximated by several models and conceptualizations. In the British context, historical thinking has been understood as a collection of practices and premises that "structure the discipline", such as "an understanding of the uses and limitations of various primary sources as evidence in reconstructing the past, and an understanding of cause and consequence, continuity



and change and similarity and difference in historical explanation” (Seixas, 2017, p. 594; but see also Bruner, 1960, and Shemilt, 1980). Based on this understanding of historical thinking, a catalogue of *key concepts* and *procedural ideas* was compiled for teachers to consider in their teaching and their assessment of historical thinking, which later became known as *second order concepts* (Seixas, 2017).

In the American context, notions of historical thinking have been less focused on understanding central concepts relevant for *doing history* but on *doing history* itself, emphasizing the importance of *reading historically* (Seixas, 2017). According to Wineburg (1991) and his student Reisman (2012), historical thinking stems from reading and working with primary sources, which involves three main steps, namely:

- *Sourcing*; identifying the type of source and its origin
- *Contextualization*: reading the document at the backdrop of the historical circumstances of its production
- *Corroborating*: comparing and contrasting the source to other sources available

The Canadian model of historical thinking, titled *The Historical Thinking Project* (Seixas, 2017; Seixas & Morton, 2013), combines the notion of second order concepts of the British framework with the pragmatic approach of the American conceptualization. Seixas and Morton (2013) defined six central historical thinking concepts called the *Big Six*, namely *historical significance*, *primary source evidence*, *continuity and change*, *cause and consequence*, *historical perspective-taking* and *the ethical dimension*. According to Seixas (2017), these concepts structure the discipline and are procedural in nature, i.e., requiring the learners to comprehend, negotiate, explain, evaluate or accommodate like in the British framework, but they put a stronger emphasis on *doing history* similarly to the American model.

In contrast to the previous approaches to historical thinking, the German contribution to the conceptualization of historical thinking is much more philosophical. Here, Rüsen’s (1983) *Theory of Historiography* has been influential, stressing the role of discipline-specific stipulations and conceptualizing *doing history* as *historical sense-making* and historical reflections (Megill, 1994; Seixas, 2017). As such, his theory identifies two tasks of historiography, namely the structuring and systematization of historical reflection as well as theorizing and examining the impact of these reflections (Rüsen, 1983). These reflections and thinking processes require discipline-appropriate ways of reasoning and self-reflection, going beyond unsystematic assumptions about the past (Rüsen, 1983), which corresponds to Gautschi’s (2015) claim that self-reflection functions as driving force for historical learning. Rüsen (1983, 2004) argues that looking back into the past embodied by historical narration helps humans orientate in the present and plan for the future both on an individual and societal level. According to Rüsen (2004),

[h]istory is a meaningful nexus between past, present, and future – not merely a perspective on what has been, *wie es eigentlich gewesen [ist]*. It is a translation of past into present, an interpretation of past actuality via a conception of temporal change that encompasses past, present, and the expectation of future events. This conception moulds moral values into a ‘body of time’. (p. 64)

As such, historical thinking is also closely connected to reflecting on and conceptualizing moral values in their historical development and contingency. Thus, for Rüsen (1983, 2004), historical thinking constitutes and articulates our historical consciousness, a central notion outlined in the following section.

So, while Rüsen's conceptualization of historical thinking might be more theoretically rich, its translation into pedagogical practice is not straightforward (Seixas, 2017), and "[c]onvincing empirical evidence [...] is still lacking" (Kölbl & Konrad, 2015, p. 19). The Anglo-American approaches, on the other hand, are designed to be much more tangible for practitioners and transformable into classroom practice but have been accused of lacking theoretical foundations and interconnections (Seixas, 2017).

#### 4.1.3 Historical consciousness

As mentioned above, Rüsen (1983, 2004) argued that thinking historically, i.e., connecting the past, present, and future meaningfully, constitutes and shapes our *historical consciousness*. Put simply, historical consciousness can be described as perceiving, experiencing, and reflecting *time*. Our historical consciousness, in turn, helps us make sense of our own existence, guides our present and future actions, and "transforms moral values into temporal wholes" (Rüsen, 2004, p. 67), mediating values and experiences through the lens of time and thus forming our historical identity (see also Rüsen, 1983). Rüsen (1983) further explains that historical consciousness is embodied by historical narration, which should be considered to be a fundamental operation of the consciousness and not 'just' an act of speech.

According to Kölbl and Konrad (2015), Rüsen's conceptualization of historical consciousness is functional in nature and can be specified on the basis of five different descriptors, those being (1) the degree of consciousness, (2) the dimension of consciousness (aesthetic, cognitive, political, rhetorical), (3) modes of articulation, (4) topoi present, and (5) the types of narrative construction (*traditional, exemplary, critical, and genetic*). The four types of narrative construction received much attention by the research community but also by Rüsen himself due to the constituting nature of narratives for one's historical consciousness (Kölbl & Konrad, 2015; Rüsen, 1983, 2004):

1. In the traditional type, the past is considered as a collection of events and their interpretations, having a direct effect on our present experience and paying little attention to the historical perspective.
2. The exemplary type identifies exemplary historical phenomena, investigates whether they are suitable to formulate universal laws, and if so, applies them to the present.
3. The critical narrative construction juxtaposes evidence and counterevidence, narratives and counter-narratives and thereby confronts moral values and historical evidence.
4. Finally, the genetic type of narrative construction focuses on change and acknowledges that historical change is inevitable and constant, temporalizing moral values and dynamizing historical identity.

Pandel (1987) approaches historical consciousness somewhat differently, focusing on mental structures, and defines seven interrelated types of consciousness, namely *consciousness of time, reality, historicity, identity, politics, economy-society, and morality* (see also Kölbl & Konrad, 2015). Another structural-analytical approach was provided by Jeismann (1980), distinguishing three

dimensions of historical consciousness (*analysis, factual judgement, value judgement*) on two axes; a semantic category (*historical interest, conception of history, historical understanding*) and an axis relating to the temporal horizon (*interpretation of the past, experience of the present, expectations of the future*). Overall, Schönemann (2012) explains that Jeisman's (1980) matrix and his understanding of historical consciousness is more narrow, focusing on the individual and their ability to reflect rather than on collective notions of historical consciousness.

As a psychological concept, "historical consciousness is understood as a mental structure or competence that underlies our dealing with collectively important aspects of past, present, and future" (Kölbl & Konrad, 2015, p. 21). Being conceptualized as a competence, this understanding highlights the performative nature of historical consciousness, shaping and reshaping one's historical identity and basing our daily acts on our interpretations of the past (Kölbl & Konrad, 2015). While there seem to be many different approaches to describing and systematizing historical consciousness, they all share that history is not understood as a review of the past but as a way of conceptualizing the present and upcoming future as a product of the past, reflecting these processes and thereby reshaping our understanding of the world and ourselves.

#### 4.1.4 Historical literacy

To begin with, historians tend to use the term *historical literacy* as a (seemingly less popular) synonym for historical consciousness (Lee, 2017; van Drie & van Boxtel, 2008). However, Lee (2017) explains that historical literacy is more closely connected to history education and is thus narrower in scope than historical consciousness, which ultimately would be independent of schooling. In any case, historians working with the notion of historical literacy, such as Lee (2017) or Maposa and Wassermann (2009), do not seem to be concerned with the linguistic realizations and languaging of historical thought and thus do not refer to linguists' works on historical literacy (see chapter 3 and especially section 3.2.2). Lee (2017), for instance, mentions that "the term 'historical literacy' begs clarification about how far it involves writing as well as reading knowledge" (p. 60), but he ultimately argues that historical literacy is not the ability to speak and write like a historian but "a historical perspective through which the world is interpreted and reinterpreted" (p. 60), completely excluding linguistic aspects of historical thinking, reading, and constructing historical narratives from their notion of historical literacy. In contrast, Maposa and Wassermann (2009) do acknowledge the role of language in their conceptualization of historical literacy, yet they remain vague and do not elaborate.

#### 4.1.5 Historical reasoning

Similarly to the psychological understanding of historical consciousness, the notion of *historical reasoning* stresses the active role of learners, adapting a sociocultural and socio-constructivist theory of learning (van Drie & van Boxtel, 2008). According to van Drie and van Boxtel (2008), historical reasoning emphasizes that learners actively use their historical knowledge to interpret past events and relate these insights to the present, much more than notions of historical consciousness or historical thinking would imply. Criticizing that most previous

conceptualizations of historical reasoning only focus on one aspect of this process, such as using evidence or explaining cause-effect, van Drie and van Boxtel (2008) propose a comprehensive framework of historical reasoning consisting of six components, those being (1) *asking historical questions*, (2) *use of sources*, (3) *contextualization*, (4) *argumentation*, (5) *use of substantive concepts*, and (6) *use of meta-concepts*. As such, this framework combines elements of Wineburg's (1991) work on historical reading (2 & 3) and the notion of second order concepts (6, *meta-concepts*) mentioned above. In addition to these, van Drie and van Boxtel (2008) added to their model *historical questions* as starting point as well as *substantive concepts* (i.e., understanding and using subject-specific terms such as *feudalism* or *Enlightenment*) and *argumentation*. Again, despite conceptual closeness, van Drie and van Boxtel (2008) do not discuss works on the language of history and historical literacy as outlined in chapter 3. However, for each of the components, van Drie and van Boxtel (2008) review relevant empirical studies and, based on their literature review, identify the main challenges for learners of history in relation to their concept of historical reasoning: First of all, learners tend to argue one-sidedly, struggling with including or balancing counter-arguments. Secondly, students rarely work with sources and when they do, they take their trustworthiness for granted and do not corroborate sources. Moreover, learners frequently do not relate a source to its historical context accurately or not at all, treating and evaluating the source like a product of the present. Such a present-day bias is also known as *presentism* and has been observed in other studies too (e.g., Carretero & van Alphen, 2014). Van Drie and van Boxtel (2008) further report that learners struggle with multi-factorial explanations and overestimate the role of individuals while underestimating the role of institutional and collective factors. Finally, the authors report that students often misunderstand and/ or wrongly apply substantive concepts.

## 4.2 The FUER Model

The notions above have seen several attempts at operationalization for pedagogical practice, resulting in a number of models of historical competency. Examples include the national standards by the *National Center for History in the Schools* [NCHS] (1996), the educational standards proposed by the *Verband der Geschichtslehrer Deutschlands* [union of German history teachers] (2006), the *FUER historical consciousness* model (Körber et al., 2007), Pandel's (2012) model of historical competences, Gautschi's (2015) model of *narrative competence*, and many more (see Barricelli et al., 2012, or Heil, 2012, for more examples and more information). Many of these models have been criticized for being too vague or not tangible enough for classroom implementation while others have been accused of being atheoretical (see Barricelli et al., 2012; Heil, 2012; Kölbl & Konrad, 2015; Seixas, 2017). In the German-speaking world, the competency model proposed by the FUER<sup>17</sup> group has been influential. More specifically, this model informed the revision of Austrian history curricula and final exam specifications and is thus reflected in

---

<sup>17</sup> *FUER Geschichtsbewusstsein* = Förderung und Entwicklung eines reflektierten Geschichtsbewusstseins [= promotion and development of a reflective historical consciousness]

history textbooks and featured in pre-service teacher education (Austrian Federal Ministry for Education, 2011, 2013; Kühberger, 2019; Kühberger & Windischbauer, 2012). As such, it is also the competency model relevant for the study at hand and will thus be outlined and discussed in the following.

Basing their model on Weinert's (2001) understanding of competence, the FUEP model conceptualizes competences as problem-solving skills that are dependent on the learners' motivational, cognitive, and social dispositions but also on their willingness to apply their skills flexibly and responsibly (Körber, 2007b). This aligns with the model's overarching objective, namely ensuring that secondary learners of history mature into responsible, reflective, and historically conscious citizens (Schreiber et al., 2007). In contrast to previous knowledge-based conceptualizations of history education, the authors of this model emphasize the importance of historical thinking and summarize their approach under the slogan: "Geschichte denken statt Pauken" ["thinking history instead of swotting"] (Schreiber, n.d.). In terms of theoretical background, the model is based on Rüsen's (1983) *Theory of Historiography* with its strong focus on reflectivity, the connection between past, present, and future, and the resulting impact on historical consciousness. As part of the FUEP research project, Hasberg and Körber (2003) developed Rüsen's conceptualization of historical consciousness into a dynamic and cyclical process model of historical thinking consisting of discrete steps.

According to this model, feelings of uncertainty or upset in the present induce the process of historical thinking by making people review and reassess assumptions, concepts, and judgements about the past, thereby revising their (personal) historical narratives (Schreiber et al., 2007). This also involves deconstructing existing narratives, revealing inherent and implied beliefs, judgements, and assumptions, which should lead to new insights, which again ideally provoke new questions and thus a new cycle of historical thought (Schreiber et al., 2007). From this spiral-progressive thought process, one can deduce three discrete steps, which constitute three of the four competences of the FUEP model, namely *Fragekompetenz* [questioning competence], *Methodenkompetenz* [methodological competence], and *Orientierungskompetenz* [orientation competence]. For the whole process of historical thinking, a fourth competence is needed, which is called *Sachkompetenz* [factual competence]. The process of historical thought is illustrated in Figure 3 on the next page, and the four competences are outlined in the following sections.

#### 4.2.1 Questioning competence

Schreiber (2007a), project leader of the FUEP group, explains *questioning competence* in the following way: Assuming that historical questions are the basic condition for and point of departure of historical thinking, history then, in its essence, is always just an answer to a question. These questions that we ask stem from uncertainty in our present experience and are thus affected by contemporary individual or collective interests, perceptions, and beliefs. Moreover, individual questions are selective and can only cover a limited area and are therefore unable to

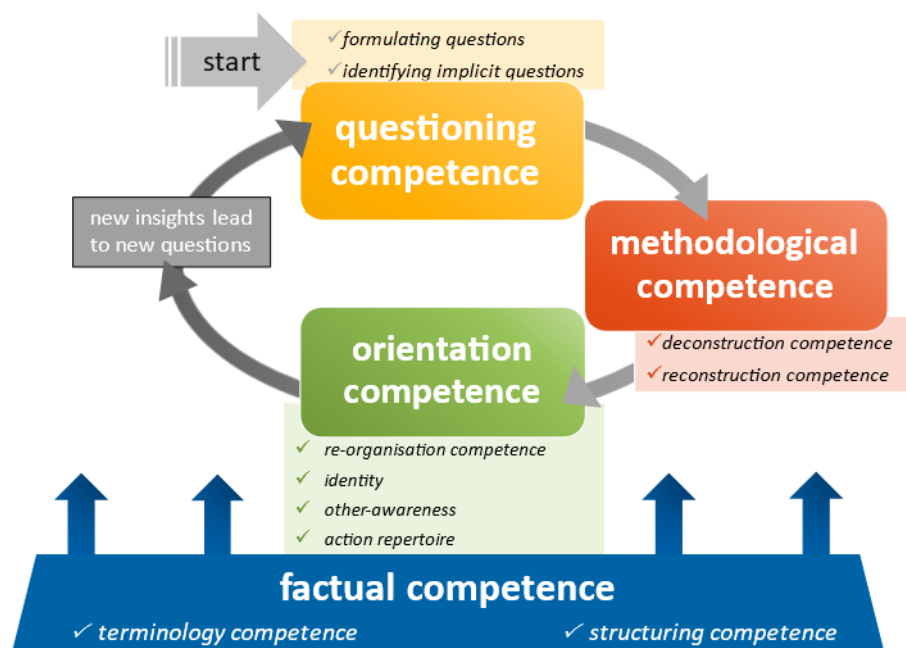


Figure 3. Representation of the FUER model based on Körber et al. (2007) and slightly modified from Dalton-Puffer and Bauer-Marschallinger (2019, p. 37)

construct a comprehensive and universal representation of the past. The FUER team (Schreiber, 2007a) describes this as *particularity principle*, which in turn consists of the *partiality*, *selectivity*, and *retrospectivity* principle. As such, this competence area involves the awareness that history does not equal objective historical recounts but is characterized by subjective and contextually-bound reconstructions. According to the FUER team, this competence consists of two core competences, namely (1) the ability to ask questions oneself and being aware that the type of question determines the outcome (i.e., the historical narrative) and (2) the ability to detect, understand, and categorize questions underlying existing narratives and to relate those external questions to one's own questions (Schreiber, 2007a).

#### 4.2.2 Methodological competence

*Methodological competence* is described in detail in Schreiber (2007b): The FUER team understands *methodological competence* as a set of skills related to working with historical sources. In the context of the historical thinking process, this constitutes the step of finding answers to the historical questions we ask. To accomplish this, historically thinking individuals first select and deconstruct a number of historical sources, which includes critical analysis of the source and its components and content in relation to its origin and historical context. On a textual basis, this might involve examining lexical choices, central terms, use of tenses, theme structure, segmentation, aspects of genre, intertextuality, and so on. On a physical level, one might investigate the materials, the colours, the condition, etc. This allows the historian to make sense of historical artefacts and assess the authenticity of a source and its content. By doing so, Schreiber (2007b) argues, students learn to read between the lines and become aware of the limitations of historical sources. A complete evaluation of validity, however, also involves a comparison to other sources, enabling the historian to assess the validity of sources and extract pieces of historical

information called *particles of the past* ([*Vergangenheitspartikel*], Schreiber, 2007b, p. 197). The set of skills related to these tasks are summarized as *Dekonstruktionskompetenz* [*deconstruction competence*], which constitutes a sub-competence of methodological competence.

To be able to provide an answer to the historical question initiating the thought process, one then needs to select and assemble the particles of the past (resulting from deconstructing sources) into a new narrative. This process is called *reconstruction* and the relevant skills are summarized as *Rekonstruktionskompetenz* [*reconstruction competence*]. This process also involves knowledge of and control over different historical genres, including sensitivity towards varying target audiences, to allow learners to create narratives in a discipline-appropriate way.

Actions of de- and reconstruction competence are assumed to be fundamental operations by the FUEP team and thus crucial for the development of historical consciousness (Schreiber, 2007b). Kühberger (2015) further argues that the process of deconstructing sources and reconstructing their own narratives enables learners to realize that sources can only help us approach the past, not capture the past in its entirety, and that historical narratives are products of their time rather than neutral representations. Such an awareness concerning sources of the past can also be transferred to narratives of the present, thus empowering learners to question given narratives (Schreiber, 2007b), e.g., deconstructing political messages or even *alternative facts*.

#### 4.2.3 Orientation competence

*Orientation competence* describes the willingness and skills to use our reconstructed historical narratives, i.e., the answers to our historical questions, to orientate in time, to adapt our beliefs and perceptions of the world and ourselves, and to inform our daily and future acts and decisions (Schreiber, 2007c). Here, the FUEP model defines four sub-competences (Schreiber, 2007c):

1. The first sub-competence is termed *Re-Organisationskompetenz* [*reorganisation competence*] and relates to the learner's readiness and ability to reflect on and acknowledge newly gained historical insights and to allow them to change their historical consciousness. Moreover, this sub-competence includes the learner's willingness to continue asking new historical questions and thereby keeps alive the process of historical thinking and consequently the process of reorganisation.
2. The second sub-competence is called *Welt- und Fremdverstehen* [*other-awareness*]. Learners competent in this area are able and willing to take over other people's perspectives, experience alterity, and revise their judgements and assumptions about others and the world more generally according to new insights. As Lamsfuß-Schenk (2010) argues, taking over other perspectives, naturally, challenges one's perception and assumptions about culture, and therefore historical learning that enables other-awareness can also be considered to be intercultural learning. Yet, Lamsfuß-Schenk (2010) adds that using the term *culture* in this respect has been controversial due to different definitions of the word. An essentialist, nationalistic conception of culture, for

instance, could foster a dichotomous differentiation between one's own culture and *the other* (Lamsfuß-Schenk, 2010). Therefore, Lamsfuß-Schenk (2010) suggests using a constructivist definition of culture that allows for multiperspectivity, such as the definition by Geertz (1993), which assumes that culture is a system of meanings that individuals of a group construct.

3. Learners should not only cultivate the willingness and ability to adjust their perception of others according to new historical insights but also their own historical identity. This sub-competence is called *Identität [identity]*. According to Seixas (2017), “democratic states have seen a tug-of-war between political demands to use school history to promote national solidarity and a liberal educational vision of history to promote an engaged, literate, critical citizenry” (p. 293) for roughly a hundred years. The sub-competence *identity* clearly shows that the FUER model favours critical, self-reflective identity construction over the instruction of master narratives.
4. The last sub-competence is called *Handlungsrepertoire [action repertoire]* and describes the willingness and ability to develop, reflect, and potentially revise tools and action-scripts to tackle current and future problems. Schreiber (2007c) adds that working on this competence should not be understood as handing them a toolkit or a manual but as a way of providing space and guidance for learners to construct, enlarge, reflect, evaluate, and revise their own repertoire of historical actions and understand their historical contingency.

To summarize, learners' orientation competence is all about the construction of meaning and making sense of history in relation to their own world of experience. As such, orientation competence can also be described as the ability to manage one's own historical consciousness.

#### 4.2.4 Factual competence

Unlike the other competences of the FUER model, this fourth competence does not form a concrete step in the process of historical thought. Instead *factual competence*, might be required throughout as it refers to a learner's willingness and ability to use and understand historical terms, concepts, categories, principles, and scripts structuring the historical domain (Schöner, 2007). In other words, factual competence describes the ability to make use of declarative, conceptual, and procedural historical knowledge (Schöner, 2007). Here, the FUER team distinguishes between two sub-competences (Schöner, 2007):

1. *Begriffskompetenz [terminology competence]* refers to one's capability of understanding, using, and relating historical terms, allowing historians to talk about history in a precise and subject-appropriate way. Drawing on Ogden and Richards' (1923) elaboration of Saussure's theory of signs, the FUER team stresses that form and meaning (*significant* and *signifié*) are tied by conventions. Such conventions are subject to historical and cultural



circumstance, and therefore being competent in this area includes understanding that meanings of terms can change over time. This also entails the ability to tell apart present and past language as well as everyday language and subject-specific language. In general, high levels of this competence are characterized by the ability to constantly enlarge, reorganize, and improve one's historical lexicon. Unlike the other (sub-)competences, terminology competence seems to be clearly concerned with language.

2. *Strukturierungskompetenz [structuring competence]* describes the ability and readiness to structure historical insights and content via the understanding and application of (a) methodological scripts, (b) fundamental epistemological principles, (c) content-related categories, (d) and subjective conceptions. For instance, being competent in this area means knowing the relevant action-patterns of deconstructing a source in order to structure historical content (→ a). Other examples include understanding and applying the principle of *particularity* or *retro-perspectivity* for the purpose of structuring historical content (→ b), the systematizing function of historical categories, e.g., *epochs* or *East-West division* (→ c), or identity-relevant conceptions, such as *nation* or *us-against-them* dichotomies (→ d).

In summary, factual competence relates to the management and application of historical knowledge. However, Kühberger (2015) adds that factual competence is not to be confused with knowledge about particular historical events or people, such as year dates or names of emperors, which members of the FUEr group termed *Arbeitswissen [working knowledge]*. This latter type of knowledge is relatively stable and established within the history community but fairly fluctuating within the memory of an individual. Therefore, Kühberger (2015) stresses the importance of flexibly activating and expanding working knowledge during competency-based history lessons.

#### 4.2.5 Grading Matrix and performative verbs

As most educational systems define educational standards and normative learning outcomes, the FUEr team added another dimension to their model called *Graduierungslogik [grading logic]*, defining three fundamental levels of competency. These levels are based on varying degrees of reflectivity, learner autonomy, and control over subject-specific criteria and conventions as outlined in Körber (2007a):

1. *Basic (or a-conventional) level:* At this level, competences are rudimentarily developed, which means that learners at this level cannot apply their skills systematically and conventionally, i.e., in a subject-appropriate way. In other words, at this stage, historical thinking happens intuitively and spontaneously without consideration of subject-specific criteria. Moreover, these learners present low degrees of self-reflectivity and autonomy.
2. *Intermediate (or conventional) level:* Learners at this stage adhere to the conventions of the discipline and apply their skills fairly systematically and autonomously, presenting their own thoughts rather than echoing those of others.

3. Elaborate (or trans-conventional) level: At this level, learners demonstrate understanding of conventional scripts, operations, categories, and concepts and apply them self-reflectively, systematically, self-reliantly, and critically. Furthermore, learners at this stage are able to adapt these scripts, operations, categories, and concepts for their own purposes if necessary.

Additionally, Körber (2007a) explains that a zero and a maximum level indicate the theoretical beginning and end of the spectrum (i.e., complete lack of competence and complete control of all competences in all situations). In between these five levels (including the zero and the maximum level), there is, theoretically, an indefinite number of fine-grained nuances of competency, indicating that learners progress continuously rather than by leaps.

Reflecting this three-staged grading matrix, history educationalists in the German-speaking world have defined three levels of history tasks, namely *reproduction* (level 1), *transfer/ reorganisation* (level 2), and *problem-solving/ reflection* (level 3) (e.g., Austrian Federal Ministry for Education, 2011; State Institute of Bavaria, 2005). To increase the operationalizability of these levels for both educational practice and assessment, Kühberger (2011) created a matrix of performative verbs, termed *Operatoren*, which function as verbs prompting pre-defined verbal actions. As such, Kühberger (2011) argues, they can facilitate the communication between learners and teachers and increase transparency of task demands. Drawing on the work by Bloom et al. (1956) and L. W. Anderson and Krathwohl (2001), Kühberger (2011) identified a list of performative verbs for each level to be used for final examinations [*Matura*] and, consequently, for task design more generally. On level 1, one would find verbs eliciting the reproduction of historical knowledge (e.g., “describe”, “list”, “name”, “summarize”, Kühberger, 2011, p. 17)<sup>18</sup>, whereas verbs on level 2 should prompt actions that involve self-reliant examination of historical input and transfer of familiar procedural knowledge to new contexts (e.g., “explain”, “analyse”, “classify”, “apply”, p. 18). The verbs on level 3 are supposed to trigger actions that require reflectivity, critical analysis, and problem-solving skills, such as “evaluate”, “interpret”, “justify”, “develop”, or “deconstruct” (Kühberger, 2011, p. 18). This matrix contains 16 different types, and for each, Kühberger (2011) provided an assumed underlying communicative intention, similar to the CDF construct. Yet, the given intentions rather function as specifications of the individual performative verbs and often do not provide the quintessence of the operation or a communicative intention per se. For instance, the communicative intention of “erklären” [explain], on level 2, is specified as relating issues and materials based on prior knowledge and insights in a justified way. Such an operation might be closer to “einordnen” [classify], which is listed as a separate performative verb. Similarly, “herausarbeiten” [explicate something] and “ermitteln” [find out] both share the underlying intention of discerning facts and relations from material provided, and therefore the State Institute of Bavaria (2005), for instance, rightfully regards these verbs as two examples of the

---

<sup>18</sup> The “Operatorenmatrix” was published in German. All translations are my own.

same type. Examples like these indicate substantial overlap and conceptual imprecision, and thus this matrix might be difficult to implement.

#### **4.2.6 Empirical validation and criticism of the FUER model**

While the FUER group describes their model and its underlying theory in great detail (over roughly 800 pages), there seems to be only limited empirical validation of their model (Dalton-Puffer & Bauer-Marschallinger, 2019; Heil, 2012; Körber & Meyer-Hamme, 2015). Those few studies dealing with the model in practice focus on textbooks (e.g., Schreiber et al., 2013) or the model's applicability for rating learner performances (e.g., Meyer-Hamme, 2007; Trautwein et al., 2017; van Borries, 2007). What seems to be absent are classroom-based studies, investigating these competences in action as well as the views of teachers and students. Yet, this appears not just to be the case with studies on the FUER model but research on history didactics in general. Especially in the Austrian context, research has focused on theorizing historical learning and textbook analysis, potentially, as Kühberger (2019) argues, because this reflects history as a discipline much more closely than empirical research. Coming back to the FUER model, there is, to my knowledge, only one study investigating these competences based on classroom data, namely my Master's thesis (Bauer-Marschallinger, 2016).<sup>19</sup> In this study, all competences could be observed in lower and upper secondary history CLIL lessons but were not equally prominent. Questioning competence, for instance, was hardly present at all, whereas reconstruction competence formed a central element at both levels. Looking at the learners' written performances testing all four competences, one could observe that, if prompted, learners could indeed perform these competences (in English even). Yet, some of the competences, such as questioning competence or action-repertoire, were difficult to elicit and thus to test. What this study has also shown is that these competences are realized via cognitive discourse functions and that there are certain patterns at play. In other words, different competences tend to be performed via certain sets of CDFs, which reflects the results of a hermeneutic analysis mapping the CDF construct onto the FUER model. In the context of pre-service teacher education, Deschner et al. (2010), investigated student-teachers' beliefs and reflections concerning competency-based CLIL history education as envisaged by the FUER team, with the future aim of creating a tertiary course that fully integrates competency-based history and CLIL didactics. Using learner diaries for data collection, Deschner et al. (2010) found that most students struggle with conceptualizing both CLIL and content perspectives at the same time. The authors therefore conclude that future didactics courses should focus on the relations and boundaries of the FUER competency model and CLIL didactics. Unfortunately, no further publications about the development of these envisioned tertiary courses could be found that would specify how these different didactics indeed could be integrated in a teacher training seminar.

This lack of empirical validation has also been one of the main points of criticism of the FUER model and, by extension, of most competency models (e.g., Barricelli et al., 2012; Heil, 2012;

---

<sup>19</sup> Parts of this study have also been published in Dalton-Puffer and Bauer-Marschallinger (2019).

Staschen-Dielmann, 2010). However, the FUEP model stands out by providing a graduation of levels (Barricelli et al., 2012; Kölbl & Konrad, 2015) and its grading logic, which has been found to be theoretically stringent (e.g., Staschen-Dielmann, 2010). However, Staschen-Dielmann (2010) and Heil (2012) criticize that there are no concrete educational standards or descriptors for practitioners and policymakers to work with. Furthermore, Heil (2012) points out that the FUEP model does not define performative verbs for the different levels. As described above, this need was addressed by Kühberger (2011). While this matrix does flesh out the linguistic dimension of the FUEP model and its grading logic, it is conceptually vague, making no reference to any work rooted in linguistics, and with its 16 performatives verbs spread over three levels and exemplified by over 45 example verbs, it is quite extensive and thus impractical for classroom use.

### 4.3 History education in Austria

The following subchapter provides an overview of the educational context this study is set in. First, the Austrian school system is briefly summarized (4.3.1), then the student body of vocational higher schools and secondary colleges of business administration more specifically is described (4.3.2). Finally, policies and guidelines for history teaching in this school type are reviewed in section 4.3.3.

#### 4.3.1 The Austrian school system

The Austrian school system is presented in Figure 4 on the following page. After one to three years of Kindergarten, Austrians typically enter formal education at age 6. After four years of primary school, learners need to choose between *academic secondary school (AHS, lower cycle)* and *middle school (MS, presented as compulsory secondary school in the figure below)*.

The curriculum for lower secondary defines two standards, a general one and a higher *AHS standard*, with different requirements, effectively creating a selective school system already from grade 5 onwards. To attend AHS, students either need to present good grades in primary school or pass an entry exam (Austrian Federal Ministry for Education, 2019). Consequently, AHS groups tend to be academically stronger and are usually more homogenous in terms of achievement level than MS students. At MS, a variety of support measures have been put into place to allow for internal differentiation and individualisation, such as team-teaching, two different assessment scales (*AHS-standard* and *standard*), support classes (e.g., German), pupils-parents-teacher meetings, etc. (Austrian Federal Ministry for Education, 2018b). After grade 8, i.e., when students are usually 14 years old, they have a variety of options, which could be roughly differentiated into academically and vocationally oriented. The *academic secondary school (AHS, upper cycle)* takes another four years and usually focuses on a certain area (e.g., natural sciences, classical and modern languages, or social sciences). In the final year, i.e., grade 12, students have to complete oral and written final exams (*Matura*) in a number of subjects.

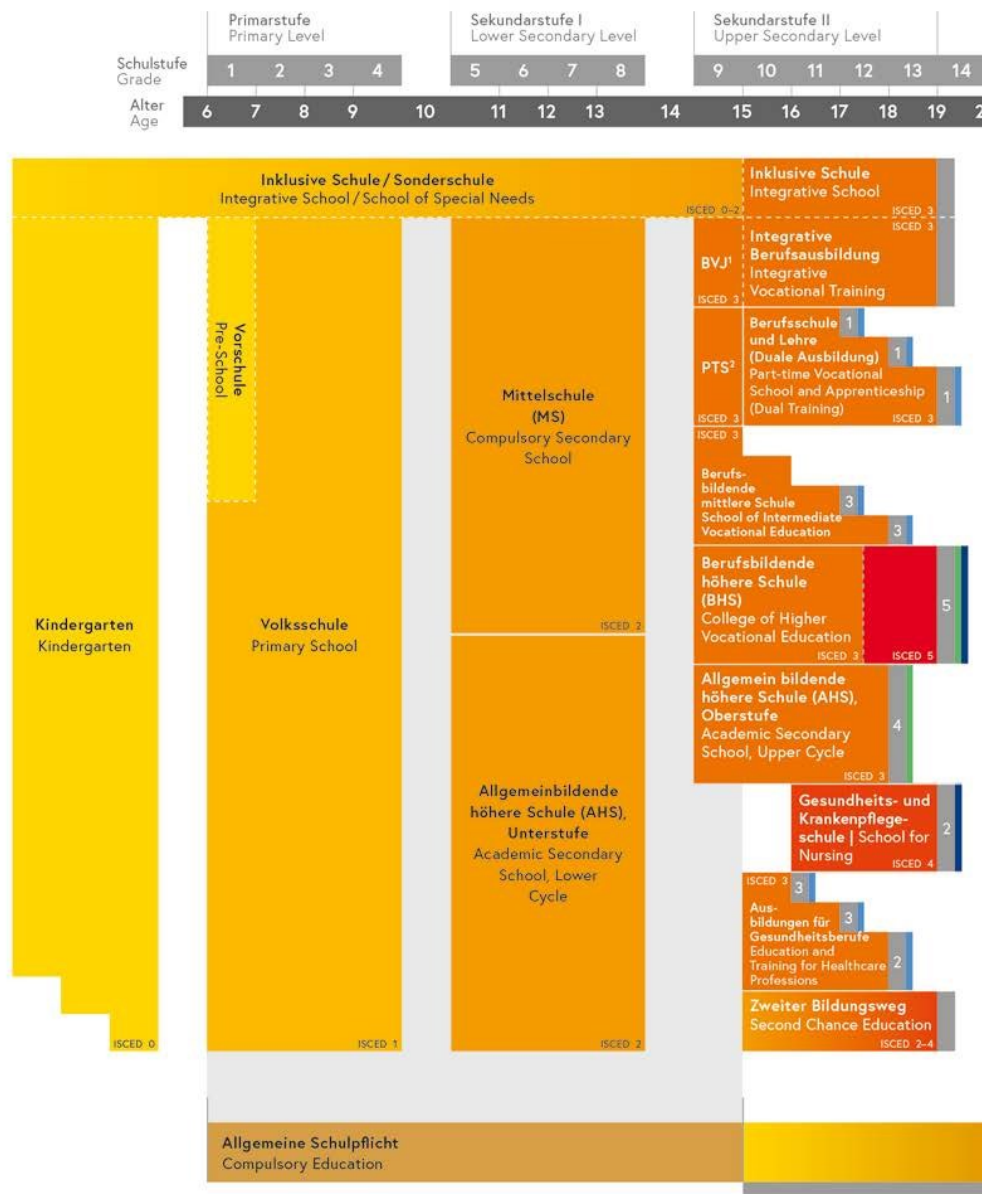


Figure 4. A schematic representation of the Austrian school system provided by the Austrian Federal Ministry for Education (2021)

Turning to vocationally-oriented schools, the shortest option would be finishing compulsory education after just one year of *pre-vocational school (PTS)*, i.e., after grade 9. This school type prepares learners for apprenticeships or other vocational post-school training. Alternatively, Austrian adolescents could attend three to four years of *intermediate vocational school (BMS)*, which prepares learners for the job market of a specific field (e.g., agriculture and forestry, different types of engineering, elementary education, fashion, tourism, trade, service and hospitality, etc.). The longest vocationally-oriented option is the *secondary college of higher vocational education (BHS)*. Like the intermediate vocational schools, these schools focus on a specific field and prepare students for entering the job market after graduation. The higher vocational schools, however, take five years and require their students to complete final exams (Matura) in a number of subjects. Both the final exams at vocational and academic schools provide graduates with university entrance qualifications. These final exams consist of a standardized

written part, a non-standardized oral part, and a pre-scientific paper (academic school) or diploma project (vocational school). The choice of subjects depends on the type of school and their focal areas, but some subjects are obligatory, such as Maths, German, and an FL, usually English. The final exams or university entrance qualification exams (*Studienberechtigungsprüfung*) can also be completed via evening schools, external programmes, and advanced training courses. Moreover, it is possible to combine apprenticeships and preparation for final exams (*dual training/ duale Ausbildung*).

#### 4.3.2 The student body of vocational upper secondary education

The school type in focus of this dissertation is the *secondary college of business administration*, i.e., a secondary school of higher vocational education with a focus on business. In terms of student numbers, in 2018/19, higher vocational schools (BHS) were the most popular choice (35.5% of all 9<sup>th</sup>-grade students), followed by upper secondary academic schools (AHS) with 28.4% of all 9<sup>th</sup>-grade students, pre-vocational schools (PTS, 19.5%), and intermediate vocational schools (BMS, 16.6%) (Statistik Austria, 2020, pp. 28–29). The majority of students at higher vocational schools (BHS) attended middle schools (MS) in lower secondary (55.1% in 2018/2019) while only 27.7% transferred from lower secondary AHS (Statistik Austria, 2020, p. 55).<sup>20</sup> In contrast, only 16.8% of students at upper secondary academic schools (AHS) completed their lower secondary education at a middle school (MS) (Statistik Austria, 2020, p. 55), indicating that the choice of school at age 10 has a considerable effect on their later educational career.

Zooming in on secondary colleges of business administration (HAK), 67.1% of all graduates (grade 13 in 2018/19) attended general or new secondary schools at lower secondary level (Statistik Austria, 2020, p. 65), indicating that most HAK learners present a less academic background when starting upper secondary. This might be reflected in a relatively high failure rate at HAK, namely 12.2% across all grades, which is the highest score of all BHS types. Those that finish secondary college of business administration, however, are more likely to enter tertiary education than any other BHS graduates, with 62.8% of all HAK graduates attending higher education within three years after graduation (Statistik Austria, 2020, p. 65, point of reference: class of 2013-2015). In comparison, at AHS, 89.2% of all graduates continue their education in the tertiary sector within three years (Statistik Austria, 2020, p. 65). In terms of gender, 56.8% of all students at commercial schools were female in 2018/2019 (Statistik Austria, 2020, p. 31).

#### 4.3.3 History education in secondary colleges of business administration (HAK)

The HAK curriculum (Austrian Federal Ministry for Education, 2014) is structured into core areas, with *political education and history* forming one subject in the core area *society and culture*. The decision to combine history and political education dates back to the introduction of civic education as a teaching principle in 1978, which has had a major effect on the conceptualization of Austrian history education (Kühberger, 2019). Moreover, in 2007, the voting age in Austria was

---

<sup>20</sup> The remaining students repeated grade 9 and/or transferred from other school types.

lowered to 16 under the condition that political education would be more present in the Austrian educational system (Kühberger, 2019). As a consequence, political education was officially integrated into the subject history in the course of curricular reforms in the past two decades (Kühberger, 2019).

In terms of hours allocated by the HAK curriculum, *political education and history* is scheduled for one weekly hour in year 2 (grade 10) and two weekly hours in year 3 and 4 (grade 11 and 12). In year 5 (grade 13), two weekly hours are allocated to the subject *international economic and cultural areas*, which also deals with topics and skills associated with contemporary history and political education. In any case, history education does not take up much space in the timetables of HAK students. For each semester, the curriculum defines one to six specific topic areas as well as three to seven educational objectives. Topics are presented in the form of key concepts, such as “[s]olidarity and exclusion” or “[s]tate and religion”, which are usually accompanied by a number of subtopics (e.g., “nationalism, racism, anti-Semitism” for the first example and “theocratical state, feudal state, secular state, fundamentalism” for the second example) (Austrian Federal Ministry for Education, 2014, pp. 92–93).<sup>21</sup> Objectives, in contrast, are specified via can-do statements, such as the following:

*The students can*

- *use historical sources critically to reconstruct and deconstruct history,*
- *present and analyze social developments and assess their importance in historical context*
- *recognize the importance of the Arts as an expression of the zeitgeist, see and critically assess artistic expressions in a historical context* (Austrian Federal Ministry for Education, 2014, pp. 92–93)

As can-do statements, these objectives centre on verbal actions. In the subjects *political education and history* and *international economic and cultural areas*, the following verbs are present in the official English translation of the curriculum: analyze (11x), apply (1x), assess (13x), assign (1x), characterize (1x), comment (1x), compare (4x), describe (6x), describe the influence(1x), defend(1x), develop (2x), discuss (5x), evaluate (2x), explain (6x), highlight(1x), identify (2x), identify motives (1x), justify (1x), name (5x), notice (the differences) (1x), present (1x), outline (1x), reason (1x), recognize(1x), reflect (3x), see causal links (1x), use sources (1x), and work out (1x) (Austrian Federal Ministry for Education, 2014, pp. 91–96). In total, there are 28 of such verbs, but they do not appear to equal degrees. Especially assessing, analysing, describing, explaining, discussing, and naming (as well as related verbs) are used frequently.

Looking into content areas, year 2 (grade 10) predominately focuses on political education and civics (e.g., “political parties and their ideological principles”, “the political system in Austria”), whereas year 3 takes a historical perspective (Austrian Federal Ministry for Education, 2014, p. 91). In year 3, the curriculum lists topics as early as the Neolithic revolution (key concept: “milestones in the historical development”) all the way up to the end of the Habsburg monarchy and the development of successor states (key concept: “conflicts between nationalities and

---

<sup>21</sup> All quotations of the curriculum are taken from its official translation (Austrian Federal Ministry for Education, 2014).

cultural conflicts”, Austrian Federal Ministry for Education, 2014, p. 92). Year 4 and 5 then focus on political, social, and economic issues of the 20<sup>th</sup> and 21<sup>st</sup> century, always with a view to their historical origins (e.g., “Europeanization and Americanization” or “[c]onflict areas in the economy, politics and society and their historic roots”, Austrian Federal Ministry for Education, 2014, pp. 93, 96). Unlike other Austrian history curricula, the HAK curriculum is not organized according to the FUEP competency model (e.g., the AHS curriculum, Austrian Federal Ministry for Education, 2004), yet its influence is apparent in the educational objectives it defines, reflecting the different steps of historical thinking and underlying principles (see subchapter 4.2).

As already mentioned in chapter 2, at secondary college of business administration, 72 hours of CLIL per year should be taught across all subjects from year 3 onwards, and therefore a number of history lessons might be realized in a CLIL setting even in schools without an explicit CLIL focus (Austrian Federal Ministry for Education, 2014). However, in my experience as a HAK teacher and as a researcher, hardly any HAK teachers are aware that a minimum of CLIL instruction has been defined in the new curriculum.

As for the final exams, HAK students can choose to include the subject *international economic and cultural areas* with a focus on history in their oral exams. As mentioned above, oral exams are not standardized nationwide. Instead, the Austrian Ministry of Education has issued guidelines for the subject teams to implement at their site, ensuring (some degree of) transparency and quality control while also taking into account school-specific focus areas (general guidelines for final history exams: Austrian Federal Ministry for Education, 2011; specifications for competency-based history education in higher vocational schools: Austrian Federal Ministry for Education, 2013). Both of these guidelines refer to the FUEP model as a basis for designing competency-based tasks and test items and provide explicit instructions, examples, and a list of performative verbs. Zooming in on the specifics of the oral final exam, the guidelines (2011) determine two content areas; one focused on source types and one listing key concepts. At each school, subject teachers need to compile a task bank comprising topics and themes from both areas, with source-based topics taking up at least one third of all tasks. Each task should consist of source materials and three to five subtasks considering all three competency levels, i.e., reproduction (level 1), reorganisation/ transfer (level 2), and problem-solving/ reflection (level 3). The source materials should not merely serve illustrative purposes but should be essential to the task to ensure a competency-based format. Furthermore, it is specified that the prompts should include performative verbs from the list provided (see Kühberger, 2011).



## **5. Methodology**

This dissertation is set within a design-based research (DBR) framework. Since this is a rather novel methodological approach, the concept of DBR is introduced and discussed on a general level in the first subchapter (5.1). Subsequently, the particular research context of this study is described (5.2). Then, an account of the research design is given, including information on participants, general organisation of the research process, steps of data collection, and ethical considerations (5.3). In the next subchapter, the individual methods of data collection and instruments are presented and reviewed in more detail (5.4). Finally, methods of data processing and analysis are described and discussed (5.5).

### **5.1 Design-based research (DBR)**

#### **5.1.1 Typical characteristics of DBR and terminology**

DBR is a methodological approach in educational research which studies “learning in context through the systematic design and study of instructional strategies and tools” (The Design-Based Research Collective, 2003, p. 5). DBR is not (primarily) concerned with “what works”, but it is interested in “how we can make something work and why” (McKenney & Reeves, 2014, p. 143). In other words, DBR “goes beyond merely designing and testing particular interventions” since these interventions are understood as embodiment of certain theoretical conjectures and claims about learning and teaching (The Design-Based Research Collective, 2003, p. 6). As such, DBR aims at improving educational practices while generating “contextually-sensitive design principles and theories” (Wang & Hannafin, 2005, p. 7).

DBR came into being as a reaction to criticism concerning the lack of practical application of educational research, taking inspiration from practically-oriented engineering and software research (Euler, 2014). Very often, scientific insights are not translated into classroom practice as practitioners simply cannot access these findings or relate them to their own experiences and classroom reality (Euler, 2014). An increasing number of researchers in the field argue that DBR “could effectively bridge the chasm between research and practice in formal education” (T. Anderson & Shattuck, 2012, p. 12). T. Anderson and Shattuck (2012) challenged this assumption by reviewing the five most cited DBR papers from each year of the previous ten years and conclude that DBR indeed seems to offer a methodology able to bridge the gap between research and practice.

The link of theory and practice inherent to DBR is a result of its dual focus: DBR is not only concerned with designing new tools and products; it is also supposed to “advance a theoretical agenda, to uncover, explore, and confirm theoretical relationships” (Barab & Squire, 2004, p. 5). Sandoval (2014) conceptualizes these focal points as commitments, namely the commitment to produce innovative opportunities for learning, the commitment to gain contextual knowledge of these learning environments, as well as the commitment to strive for foundational knowledge about teaching or learning in relation to the innovation. Sandoval (2014) adds that the

simultaneous consideration of all these commitments is unique to design research, but there seems to be no clear methodological protocol for DBR, leaving the choice and arrangement of research methods up to the researcher. Instead, most handbooks and articles on design research “largely articulate an ethos of design research, described in terms of aims or commitments of the approach”, which is often exemplified by referring to studies which embody these commitments and typical characteristics (Sandoval, 2014, p. 19). In this sense, Wang and Hannafin (2005, p. 8) outline a number of defining as well as common characteristics of design-based research, illustrated in Table 4 below:

pragmatic	<ul style="list-style-type: none"> <li>– Design-based research refines both theory and practice.</li> <li>– The value of theory is appraised by the extent to which principles inform and improve practice.</li> </ul>
grounded	<ul style="list-style-type: none"> <li>– Design is theory-driven and grounded in relevant research, theory, and practice.</li> <li>– Design is conducted in real-world settings and the design process is embedded in, and studied through, design-based research.</li> </ul>
interactive, iterative, and flexible	<ul style="list-style-type: none"> <li>– Designers are involved in the design processes and work together with participants.</li> <li>– Processes are iterative cycle of analysis, design, implementation, and redesign.</li> <li>– Initial plan is usually insufficiently detailed so that designers can make deliberate changes when necessary.</li> </ul>
integrative	<ul style="list-style-type: none"> <li>– Mixed research methods are used to maximize the credibility of ongoing research.</li> <li>– Methods vary during different phases as new needs and issues emerge and the focus of the research evolves.</li> <li>– Rigor is purposefully maintained and discipline applied appropriate to the development phase.</li> </ul>
contextual	<ul style="list-style-type: none"> <li>– The research process, research findings, and changes from the initial plan are documented.</li> <li>– Research results are connected with the design process and the setting.</li> <li>– The content and depth of generated design principles varies.</li> <li>– Guidance for applying generated principles is needed.</li> </ul>

Table 4. Characteristics of design-based research according to Wang and Hannafin (2005, p. 8)

As for terminology, while *design-based research* is used by, for example, Wang and Hannafin (2005), T. Anderson and Shattuck (2012), Barab and Squire (2004), Euler and Sloane (2014) and The Design-Based Research Collective (2003), another label commonly used is (*educational*) *design research*, e.g., by Van den Akker, Gravemeijer, McKenney and Nieveen (2006a), Sandoval (2014), Plomp and Nieveen (2013), and McKenney and Reeves (2012). McKenney and Reeves (2012) also name a number of defining features of *design research*, which are rather similar to Wang and Hannafin's (2005) list of typical characteristics of *design-based research*. According to McKenney and Reeves (2012, pp. 13–15), *design research* is

- *theoretically oriented*, as existing theory is used as a basis for scientific investigation, which is aimed at developing this theory further while also addressing practical problems.
- *interventionist* since it usually involves an intervention which should bring about positive change. In design research literature, the term *intervention* is very often used

synonymously with *innovation* and refers to any kind of solution designed to address a specific problem.

- *collaborative* in the sense that the expertise of practitioners and researchers inform each other.
- *responsively grounded* because it relies on theory, participant expertise, and empirical research in naturally occurring test beds. As such, it “is structured to explore, rather than mute, the complex realities of teaching and learning, and respond accordingly” (McKenney & Reeves, 2012, p. 15).
- *Iterative*, as it involves multiple research cycles to develop and fine-tune the interventions as well as theoretical insights.

Van den Akker, Gravenmeijer, McKenney and Nieveen (2006b), also using the term *design research*, list a number of typical characteristics and highlight five, which are again related or even identical to the ones mentioned above: *interventionist*, *iterative*, and *orientation towards process, utility, and theory* (van den Akker et al., 2006b, p. 5). They also explain that there are many different but closely related conceptualizations of this research paradigm, resulting in an abundance of labels besides *design research* and *design-based research*. Other labels would be *design studies*, *design experiments*, *development/ developmental research*, *formative research*, *formative evaluation* and *engineering research* (Euler, 2014; van den Akker et al., 2006a; Wang & Hannafin, 2005). In the spirit of pragmatism – which seems to be a key characteristic of DBR – this dissertation uses the term *design-based research* as common label referring to the whole family of related and similar approaches while pragmatically drawing on the conceptualisations mentioned in this section as seen appropriate for the research interest of this PhD project.

### 5.1.2 Methods and research design in DBR

As mentioned above, there is no clear methodological protocol for DBR, yet there are a number of methodological conventions and principles based on typical characteristics of DBR. First of all, DBR studies take place in naturally occurring test beds so that newly developed theories and tools reflect and consider real-life classroom practice (Kelly, 2006). Of course, these settings involve numerous variables as well as practical constraints. Unlike randomized field trials, design-based researchers do not intend to “randomize away” or isolate these influences (Kelly, 2006, p. 113). Instead, it is necessary to characterise them and gain a comprehensive picture of the situational context (McKenney et al., 2006). Moreover, in authentic contexts, new factors can emerge, and therefore it is crucial to create a flexible research design that consists of a number of different data collection methods (McKenney et al., 2006). Given these considerations, DBR mainly makes use of qualitative research methods. However, quantitative methods might complement the research design, and qualitative results can be quantified, allowing triangulation of methods and thereby increasing validity (Bakker & van Eerde; McKenney & Reeves, 2014).

Another key element in DBR methodology is its cyclical research process as illustrated in Figure 5. This model of a generic DBR process has been created to illustrate how the process of DBR is understood in the context of this dissertation and basically presents a synthesis of models by Fraefel (2014) and McKenney and Reeves (2012). McKenney and Reeves (2012) reviewed several

models and frameworks for educational design research, such as Baek and Bannan-Ritland (2008), Gravemeijer and Cobb (2006), Reinking and Bradley (2008), McKenney et al. (2006), or Wang and Hannafin (2005), and constructed an adaptable model. This model contains three key phases, namely *analysis/exploration*, *design/construction*, and *evaluation/reflection*, which can be flexibly and iteratively arranged. This model also describes how the implementation of the design expands over time as the intervention and theoretical understanding mature. While this model identifies key stages of DBR, the iterative and cyclical nature of this approach does not come out clearly. Furthermore, it gives the impression that classroom implementation happens continuously and constantly over time, which does not necessarily have to be the case. Fraefel's (2014) model, on the other hand, clearly visualizes the cyclicity and iterativeness of DBR and also depicts the implementation phase as happening in stages, closely connected to the designing of interventions. This model, however, neither highlights the continuous development of theory that is essential to DBR, nor does it include a number of key steps mentioned in other frequently cited DBR reference works, such as the above-mentioned model by McKenney and Reeves (2012). Looking at both these models, it seems that they can counteract each other's weaknesses, resulting in a more complete and illustrative representation of a generic DBR process. In line with McKenney and Reeves (2012), the model proposed here is conceptualized in a flexible and customizable manner to account for the high degree of variety in design-based research.

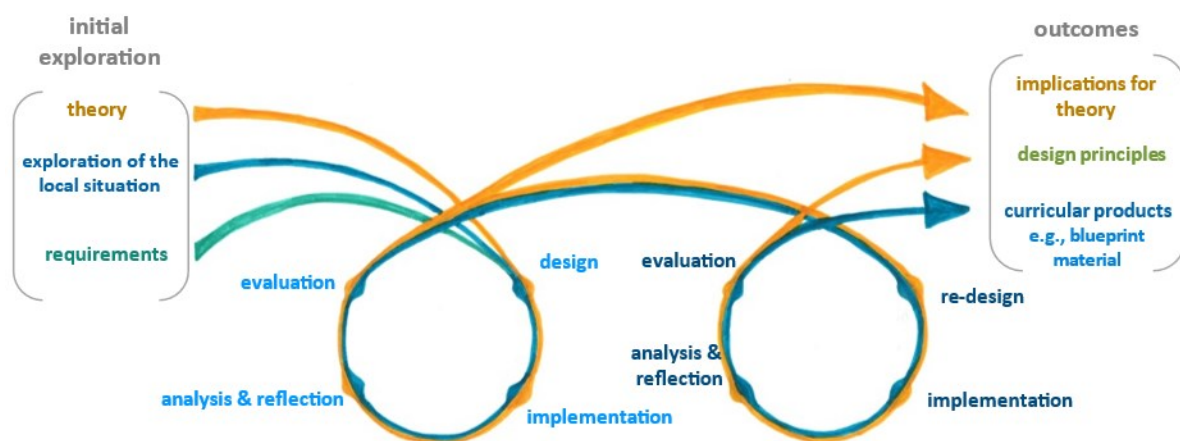


Figure 5. DBR as a cyclical research process based on Fraefel (2014) and McKenney & Reeves (2012)

The model suggested above shows that at the beginning of a DBR project, the local situation, (curricular) requirements, as well as theory should be taken into consideration (Fraefel, 2014; Kelly, 2006; McKenney & Reeves, 2012). Based on these, interventions addressing the specified problem are designed, ideally in collaboration of researcher and practitioner (McKenney & Reeves, 2012). Then, the teacher implements the newly designed tools while the researcher observes (Euler, 2014; McKenney & Reeves, 2012). As a next step, the process is analysed and reflected upon, and the intervention is formatively evaluated. In early cycles, the so-called *alpha stage*, evaluation focuses on internal logic, soundness, and viability (McKenney & Reeves, 2012). In later stages, evaluation is more concerned with how working components function in real-life contexts, which is often termed as *beta-testing* (McKenney & Reeves, 2012). At any stage, the

results of these investigations should feed into adaptations and fine-tuning of the interventions for the next cycle (McKenney & Reeves, 2012). Theoretically, one could repeat these cycles as often as deemed purposeful (T. Anderson & Shattuck, 2012). At the end of the final cycle, the tools developed are also summatively assessed, looking at effectiveness for the groups involved (Euler, 2014). In a follow-up study, it would make sense to conduct randomized field trials with the finalized intervention in order to allow statistical generalization and thereby determine large-scale effectiveness, which are then called *gamma-studies* (Euler, 2014; McKenney & Reeves, 2012).

### 5.1.3 Practical outcomes and theorizing in DBR

Concerning outcomes, in the course of DBR projects, curricular products, materials, and/ or pedagogical tools are developed, e.g., teacher guides, syllabi, teaching materials, learning software, or professional development aids (McKenney et al., 2006). As for theoretical implications, DBR studies are not concerned with theory testing but theory building and cultivating, which is also connected to the largely qualitative nature of DBR (Kelly, 2006; Shavelson et al., 2003). This entails that theory should not only be generated at the end, but theoretical abstractions should accompany the whole process, resulting in a continuously developing and specified body of theoretical implications. As for generalization, this type of theorizing obviously does not permit statistical generalization but allows for analytic generalization (McKenney & Reeves, 2012), i.e., generalization to theoretic models rather than to other populations (Firestone, 1993).

Another form of outcome, somewhat between theory and practice, are design principles, which Van den Akker (1999) defines as follows:

“If you want to design intervention X for the purpose/function Y in context Z, then you are best advised to give that intervention the characteristics A, B, and C, and to do that via procedures K, L, and M, because of arguments P, Q, and R.” (p. 5)

As illustrated by this quote, design principles do not attempt to be contextually independent (Kelly, 2006). McKenney et al. (2006) add that “design principles are not intended as recipes for success but to help others select and apply the most appropriate substantive and procedural knowledge for specific design and development tasks in their own settings” (p. 73). Within these limitations, design principles can also be considered to be prescriptive and action oriented. Yet, it is always up to the ‘user’ and their expertise to adapt them for their own setting (McKenney et al., 2006). In this sense, DBR results also allow for case-to-case generalization, i.e. transferring (parts of) pedagogical tools and theoretical implications to other contexts (McKenney & Reeves, 2012). As Brown (1992) puts it, “an effective intervention should be able to migrate from our experimental classroom to average classrooms operated by and for average students and teachers” (p. 143).

## 5.2 Context of the present study

This study takes place in two different upper secondary vocational schools (BHS) with a focus on business administration (*HAK*) in Vienna. As mentioned in section 4.3.1, BHS schools take five years, starting at grade 9 and leading students to partly standardized final exams (*Matura*) at

grade 13. As mentioned in section 4.3.1, higher vocational schools offer both general education and professional training, permitting graduates of this school type entrance to tertiary education as well as practice of certain professions legally regulated under commercial code in Austria and other EU member states (Austrian Federal Ministry for Education, 2017a).

The researcher has taught in both of these schools and is thus familiar with the context. Having some 'insider knowledge' is especially useful in a context-sensitive study (see Dörnyei, 2007; McKenney & Reeves, 2012) such as the present one.

### 5.2.1 School A

School A is a public school located in the southwest of Vienna and is attended by approximately 1600 students. This school offers a number of different branches, including intermediate (*HAS*) and higher vocational business education (*HAK*), bilingual higher vocational business education (*bilinguale HAK*), upper secondary add-on courses (*Aufbaulehrgang*), mono- and bilingual post-secondary vocational courses (*Kolleg*), as well as adult evening courses (*Abendschule*).<sup>22</sup> For admission to these programmes, certificates of previous education are checked, and prospective students are interviewed. Applicants for the bilingual programmes (*HAK* and *Kolleg*) need to have good grades in English and also have to participate in an English interview. In bilingual programmes, half of the overall class-time is required to be in English. In these branches, native speaker teachers often co-teach with subject teachers. Furthermore, some teachers are qualified subject teachers as well as native English speakers, especially in business classes. Other content subjects are mostly taught by teachers who are qualified English as well as subject teachers. In the bilingual secondary school, students can complete their work placement in an English-speaking country, and final exams are also bilingual.

On their website, the school states that they are committed to principles of CLIL teaching.<sup>23</sup> In addition to English, students are required to choose French or Spanish as a second FL and can attend extra-curricular Russian courses on a voluntary basis. In their mission statement, the school highlights the importance of internationalisation and linguistic diversity, practice-orientation, innovative technology, sustainability, health, and teamwork. They stress the usefulness of their peer-learning programme, co-operative learning, as well as their dedication to active and open learning. As for technological equipment, one PC connected to the internet and a projector are provided in each classroom. Furthermore, there are several computer rooms which can be booked in advance.

### 5.2.2 School B

School B is a charter school located in one of the outer districts in the northeast of Vienna. The institution supporting this school is a Viennese commercial fund. There are two different

---

<sup>22</sup> For further information on these school types see Austrian Federal Ministry of Education (2017) as well as section 4.3.1.

<sup>23</sup> No reference is provided so that the school and the participants cannot be identified.

programmes at this school, namely intermediate (HAS) and higher vocational education (HAK), accommodating approximately 600 students in total. Six years ago, i.e., before CLIL became anchored in the HAK curriculum, CLIL was introduced for one class per grade in the higher vocational programme (HAK). Here, 72 English CLIL lessons should take place from year 1 (grade 9) onwards. The choice of subjects and distribution among subjects are not specified. Instead, all CLIL sequences are logged into an electronic class register, and it is the class teacher's responsibility to make sure that at least 72 lessons are reached in total. To teach CLIL classes, no fixed criteria apply, but teachers either should have completed CLIL trainings or they need to prove their (language) competence otherwise. Students need to produce documentation of good grades in English and other languages to be admitted to the CLIL class. However, there are no official minimum standards and no other admission procedures for the CLIL strand. As for other languages taught in the programme, students have to choose one extra FL (Spanish, Italian, Russian, or French). This school also offers voluntary supplemental courses in Spanish, Russian, French, or Latin, as well as a course preparing students to take English Cambridge Certificate examinations (FCE/ BEC Higher). On their website, school B highlights the use of laptops and the implementation of IT aspects in all subjects. The school is equipped with four PC rooms, and projectors are installed in all classrooms. In their mission statement, they stress the importance of performance, critical thinking skills, responsibility, professional work ethic, entrepreneurship, promotion of individual talents, practice-oriented education, linguistic diversity, and language competence, in addition to respect, honesty, transparency, and well-being.

### 5.3 Research design

The present study intended to develop and continuously refine competency-based, CDF-focused history materials. This process was determined by and systematized via the following main research questions (RQ):

- RQ1: What kind of content-and-language-integrative pedagogical measures and materials (type and features) are needed to help students improve and elaborate their verbalization of cognitive processes (CDF use) as
  - (a) perceived by learners,
  - (b) reported by teachers,
  - (c) observed in written student performances?
- RQ2: How do students respond to explicit teaching of CDFs in the history CLIL classroom as
  - (a) reported by learners,
  - (b) perceived by teachers?
- RQ3: What is the effect of CDF-oriented teaching on the learners' development of historical competences and academic language skills as observed in written performances?

Typically for DBR, this study was organized in research cycles, one pilot and three main cycles as illustrated in Figure 6 on the next page.

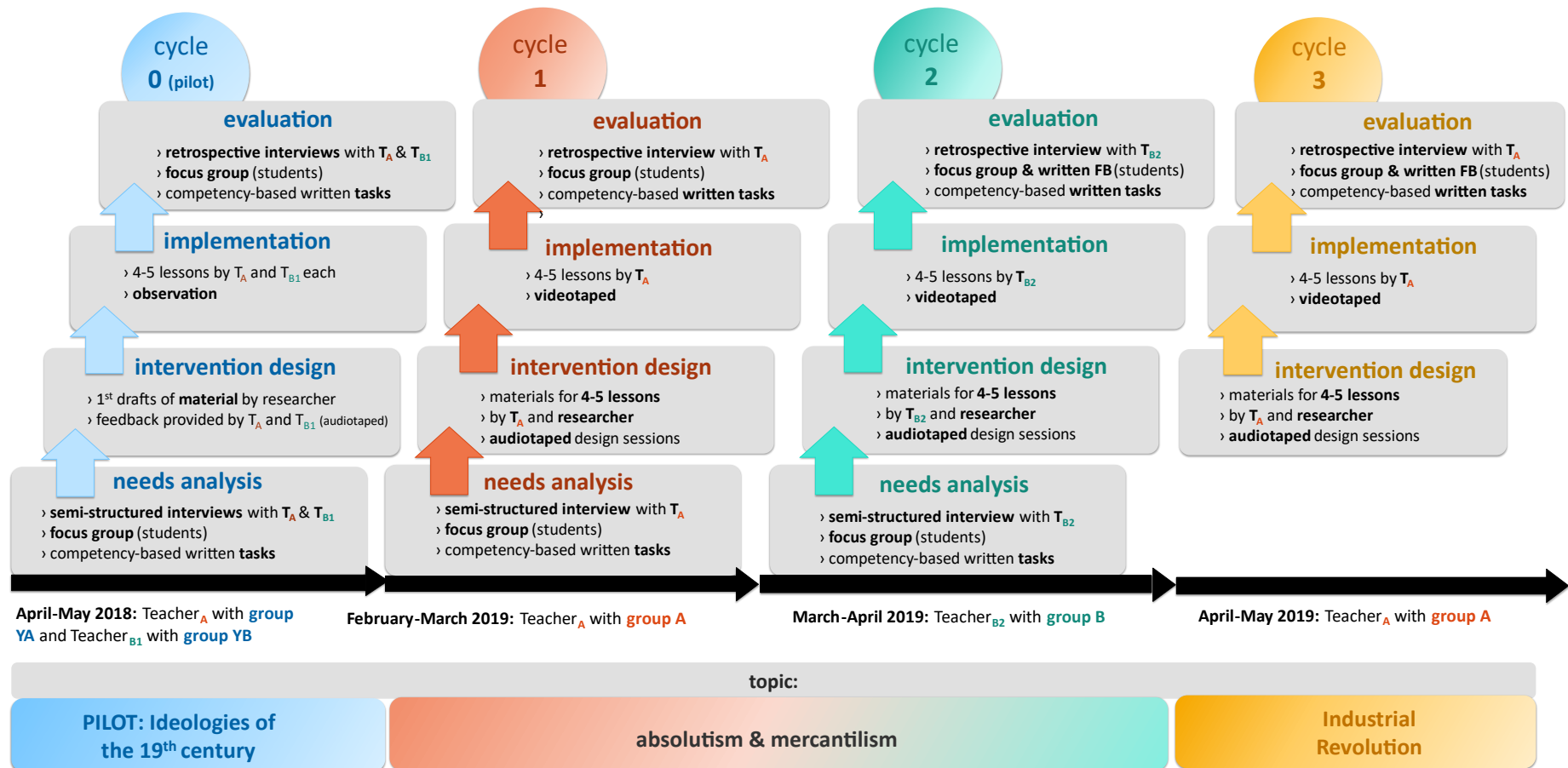


Figure 6. Overview of the research design and data collection



At the beginning of a cycle, the needs of students and teachers were identified with the help of various data collection methods, such as individual interviews, focus group interviews, and written tasks, providing insights for RQ1 (*needs analysis*). Based on these insights, pedagogical interventions incorporating CDF theory were produced by the researcher in collaborative design sessions with the teachers involved (*intervention design*). The teachers then implemented these pedagogical tools in their classroom (*implementation*). Addressing RQ2 and RQ3, the process and product were formatively evaluated using retrospective interviews with teachers (individually) and students (in groups), written tasks, and, partly, written student feedback (*evaluation*). These key steps briefly introduced here are described more thoroughly in section 5.3.3. In accordance with the results of the evaluation phase, the intervention was adapted and then re-implemented, followed by another evaluation process.

For the most part, the basic structure of these cycles stayed the same throughout the study, yet each cycle fulfilled a slightly different purpose, resulting in small differences in the research design, which are outlined in section 5.3.4, where also more information on the sequence of steps and cycles will be provided. First, however, section 5.3.1 introduces the participants since they play a crucial and process-determining role in DBR, followed by an outline of the measures taken to protect the participants' interests and privacy (section 5.3.2).

### **5.3.1 Participants**

#### **5.3.1.1 Teachers**

To effectively link theory and practice, I closely worked with teachers who are also interested in the problem addressed by this study and who were willing to actively contribute to potential solutions. As such, the practitioners' role was conceptualized as a productive participant whose practical and professional expertise as well as their realistic estimations concerning practicality of pedagogical tools would ensure the creation of interventions viable for real-life classroom implementation (see van den Akker & Nieveen, 2016). According to McKenney et al. (2006), the research process should ideally be understood and organized as a means for professional development, providing opportunities to reflect on and directly address issues relevant for teaching practice. As these tasks constitute additional expenditure of time and require commitment and willingness to collaborate, all teachers involved in this project were personally recruited from my professional network to facilitate successful collaboration and reduce the risk of any dropouts. The teachers involved were aware of the entailments of participating in this study. Furthermore, they all stated that professional development and finding solution approaches relevant for their classroom reality were driving factors for participating in this study. They were all very positive towards BE and CLIL but did feel that there was room for improvement.

Teacher A (cycle 0, 1, 3): Teacher A has been teaching English as well as history and political education at school A since 2000. She has also been teaching history and political education bilingually since then, as her willingness to do so was a prerequisite for her employment at this school. During her studies, she had not completed any kind of further education or module on CLIL or BE simply because there was nothing available.<sup>24</sup> She is, however, a qualified mentor teacher for novice teachers in both her subjects and regularly attends in-service teacher trainings on various topics. Her first language is German.

Teacher B1 (cycle 0 - pilot): Teacher B1 has been teaching German as well as history and political education since 2013. He started working at school B one year later and was asked then whether he considered himself able to teach history and political education in English, which he affirmed. Since then, he has taught CLIL classes every year. He had not completed any courses on CLIL or BE during his studies but has participated in a school-internal in-service teacher training concerning CLIL. His first language is German.

Teacher B2 (cycle 2): Teacher B2 has been teaching history and geography for 13 years and also coordinates history education at school B. When he first started studying, he was also enrolled at the English and German department, but after the first half of his studies, he only continued with history and geography because he realized that he did not want to be a language teacher. However, when CLIL was introduced at his school six years ago, he welcomed the initiative and wanted to be part of the programme. Around this time, he also developed a keen interest in language-sensitive teaching and joined a group initiated by the Ministry of Education with the purpose of designing language-sensitive history and geography materials for upper secondary vocational schools. Additionally, he has repeatedly organised in-service teacher trainings on language-sensitive teaching, which also include CLIL methods. As for further training regarding CLIL, he had never done a whole course or module but completed a number of shorter CLIL trainings. His first language is German.

### **5.3.1.2 Students**

The student groups involved in this study were the students of the participating practitioners. All of them attended grade 11 at the time of data collection in one of the two vocational schools described in subchapter 5.2 and were taught history either by T<sub>A</sub>, T<sub>B1</sub>, or T<sub>B2</sub>. Grade 11 has been chosen because this is the only grade of this school type focusing entirely on history education and not on its related fields that are part of the subject cluster *society and culture* (see section 4.3.3 for more information). Grade 11 marks the third year of upper secondary education, so most students were 16 to 17 years old, although some students were older due to class repetition or change of school type. The students participating in this study had opted for the bilingual branch in school A and the CLIL group in school B, respectively.

---

<sup>24</sup> Presently, a variety of in-service teacher trainings are available but not compulsory. See subchapter 2.2 for more information.

For this study, all students were asked to fill in a brief survey on their age, gender, first language(s), L2s, their English and history grades of the previous year, and whether they have had any BE prior to attending this programme. In general, these student groups were relatively homogenous within classes, as all students had comparable backgrounds and shared similar experiences with regards to the research focus of this dissertation.

Students YA (pilot cycle): This group attended grade 11 in school A at the time of data collection. In this group, there were 12 female students and no male students. Eleven of these students completed the survey mentioned above. Three of them reported that they had more than one first language. Three other students did not speak German as (one of) their first language(s). All students learned a third language at school, with all except one having opted for Spanish. The group achieved a 2.91 average in English and 2.0 in history the previous year, with 1 being the best grade and 5 marking 'fail'. Finally, four students reported that they had experienced some sort of English BE before enrolling at school A. According to their history and English teacher (Teacher A), this group lacked motivation concerning history education as well as language competence. T<sub>A</sub> also taught a second bilingual group at grade 11, but we agreed to include the above-described group in this study for maximum impact since we assumed that this group could benefit more from a novel approach.

Students YB (pilot cycle): Pilot group YB at school B consisted of 21 students (14 female, 7 male). Nineteen of them filled in the personal data questionnaire. Five students reported that they had two first languages. Two of these five stated that German was not one of their first languages. In total, there were ten students with first language(s) other than German, with Punjabi being the most frequent one, followed by Serbian, and Polish. Six students spoke more than three L2s. As for FLs learned at school other than English, most students chose Spanish or French. Concerning grades, the average in the previous year was 2.18 for English and 1.71 for history. Seven students had previous experience in terms of English BE. However, most of them added that the intensity of these programmes had been rather low. Furthermore, one student attended a fully-fledged German-Czech bilingual middle school.

Students A (cycle 1 and 3): Group A at school A included 15 female and 4 male students, with an age range of 16 to 20 due to four students repeating a class, including one student who had recently moved to Austria and thus lacked the required German language skills. Thirteen of them reported a different L1 than German, with Serbian and Turkish being the most frequent ones. However, three of those stated that they also used a lot of German in their everyday lives, and another student whose L1 was not German listed two first languages (Romanian and Serbian). All students learned English and a second modern language at school (Spanish or French). Their previous average English and history grades were the lowest of all groups involved in the project, with values at 3.24 (English) and 2.89 (history). Their comparatively low level of achievement was also the teacher's reason to choose this class for the project and not their parallel group in the hope to counteract their decreasing achievement levels. Finally, only three students had experienced any form of BE before attending school A.

Students B (cycle 2): This class consisted of a total of 27 students. Three students opted not to participate in the study and were thus seated where they could not be filmed. Additionally, none of their contributions in class were transcribed. Out of the 24 remaining students, 23 filled in the personal data questionnaire. Of these 23, 12 were female and 11 were male. In general, this group was very multilingual and linguistically diverse, representing 21 different languages as L1s and L2s. Four students described themselves as bilingual and one as trilingual, with German, Tagalog, and English being the most common languages amongst the multilinguals. Five others stated that their first language(s) did not include German but Tagalog, Spanish, Serbian, Arabic, or Bosnian, respectively. All students learned English and another L2, which was Spanish in most cases. Additionally, nine students mentioned a third L2, one student listed four L2s, and another student reported that he spoke seven different languages (two L1s and five L2s). Five of the 23 students reported that they had experienced some form of BE prior to attending school B. The age of the students ranged from 16 to 18 years due to three students repeating a class. The achievement level as measured by the Austrian grade system was higher than group A's, with an average grade of 2.0 in English and 1.96 in history in the previous year.

### 5.3.2 Ethical considerations

For this study, a great amount of empirical data, also specific to individuals, was collected and much of this data stemmed from adolescents who were not yet of legal age. To make sure these circumstances were fully accounted for, a number of measures were taken.

As a first step, a motion to conduct this research project was filed to the headteachers of the schools in question, considering the statutory provisions specified in the decree of the Viennese educational board, effective from December 2017 (ERIIIB: 270 §, 2017). In this decree, it was ruled by the Viennese educational board that permission of research projects is under the jurisdiction of the school in question. After preliminary approval of the headteacher, the research project in general and the specifics of data collection were presented to the statutory elected panel of teachers, parents, and pupils of the respective school, who then voted on permission. Both panels voted in favour of the project, thereby allowing data collection at their school ([appendix<sup>25</sup> section IV \(informed consent\), file B](#)).

Concurrently, an application to the Ethics Committee of the University of Vienna was submitted. The Ethics Committee of the University of Vienna evaluates research projects with or on human beings in terms of data protection and privacy, physical or psychological integrity, and rights and interests of participants (Ethics Committee of the University of Vienna, 2021). The application for this research project was positively reviewed by the Committee, confirming that all their criteria were met (see [appendix section IV, file C](#)). Then, all participants were personally informed about the study, the voluntariness of participation, data collection, privacy and data protection, and their rights in terms of withdrawal or viewing their data. Information sheets and informed consent

---

<sup>25</sup> The appendix of this thesis is available online and, in hard-copy versions, on a USB drive enclosed. For more information, see chapter *Digital appendix*.

forms (see [appendix section IV, file A](#)), which were also positively reviewed by the Ethics Committee, were handed out to all participants and, in case they were minors, to their parents too. The participants of this study all signed these forms, which are now safely stored in the researcher's office.

Methodological measures, as outlined in the application to the Ethics Committee, included the anonymization of personal data via various procedures. To begin with, all names were replaced with codes. To still be able to track the students' performances, students were given codes, consisting of the second and third letter of their first name, the first letter of their mother's name and their month of birth in digits. That way, students could reconstruct their codes in case they forgot them without realistic danger of identification. Moreover, all data that could possibly identify the participants was replaced with placeholders in the transcripts. Another strategy to protect the participants' privacy was the deletion of all raw data once the study was completed. Additionally, all electronic (and anonymized) data is protected via password and stored on non-commercial servers of the University of Vienna. Physical (anonymized) data is stored in lockable cupboards in the researcher's office, which is also locked whenever empty.

### **5.3.3 Key steps of a research cycle**

Each cycle consisted of four main stages, termed as *needs analysis*, *intervention design*, *implementation*, and *evaluation*. These steps were consecutive and built on each other. The individual steps are outlined below. Detailed information on individual methods can be found in subchapter 5.4.

#### **5.3.3.1 Needs analysis**

In the needs analysis phase, two main objectives were pursued. First, most important needs and demands in terms of CDF use and competency-oriented history learning were identified and described. Secondly, the local context was thoroughly explored since ample knowledge on contextual variables, such as school climate, resources available, student population, or system factors, are crucial in DBR or any other mainly qualitative research framework for that matter (McKenney et al., 2006). Furthermore, research has clearly shown that teacher beliefs and attitudes are a significant factor for the success of intended change and pedagogical innovation (Dijkstra et al., 2017). As for the role of students, their voices in pedagogical design have, for the most part, been neglected in educational design and CLIL research, even though their perspective has often been considered crucial for the success of a new educational design, as has been argued in chapter 2 (see, e.g., Groundwater-Smith & Mockler, 2016; Skinnari, 2020). To these ends, semi-structured interviews were conducted with teachers individually and with students in focus groups. Furthermore, students completed competency-based written tasks in the form of historical source analysis for diagnostic purposes on a topic they had previously dealt with in history class. Additionally, these tasks served as reference points to track the students' development.

### **5.3.3.2 Intervention design**

Based on the insights gained in the previous stage, curricular requirements, and CDF theory, the teacher and I collaboratively created interventions. In connection to this, McKenney et al. (2006) advise that researchers should be prepared and willing “to take on the additional role of designer, advisor, and facilitator, without losing sight of their primary role as researcher” (p. 84). This way, they argue, synergy between practice and research can be maximally exploited, provided that there is clear communication between researcher and practitioner.

For the purpose of intervention design, I repeatedly met the teachers personally to draft, develop, elaborate, and discuss lesson plans and materials for the respective unit of four to five lessons. These design sessions were audio-recorded to document our workflow, including assumptions, conjectures, and open questions. These meetings took place once or twice before each implementation phase, depending on the teacher’s schedule and the need for a follow-up meeting. Apart from meeting personally, we stayed in contact via e-mail and telephone to discuss follow-up tasks and further organisational issues. Furthermore, all materials were uploaded to an online file-share repository, which both the teacher and I could access and edit.

### **5.3.3.3 Implementation**

Then, the teacher implemented the intervention (i.e., four to five CDF-based lessons) in their own history classroom while I was observing. In the main cycles, these lessons were also videotaped to be able to document the implementation process and contextual variables, which is crucial according to McKenney et al. (2006). At this stage, the tools developed could be tried out in authentic contexts, documenting how the intervention operated in naturalistic test beds while also providing a basis for improving the intervention for future cycles or ‘real-life’ use. According to McKenney et al. (2006), these try-out phases in naturally occurring test beds are essential to draw legitimate conclusions on local viability and robustness of the design.

### **5.3.3.4 Evaluation**

At this last stage, the perspective shifted to retrospective, analysing and reflecting on the design process as well as evaluating learner products. Shortly after the implementation, the focus group was interviewed again, evaluating the intervention from the learners’ point of view. In these interviews, students were also invited to contribute further ideas for the subsequent adaption of the materials or future interventions. In later cycles, student interviews were complemented by short written feedback forms. In the case of cycle 2 (school B), these feedback sheets were handed to students who were absent on the day of the interview because of a language contest scheduled for the same day (which I had not been informed about). With the Easter break approaching, it was decided not to reschedule but to ask these students to fill in a short written open-ended feedback form (see [appendix/ section I/ E](#)) in order to gather the evaluations of students with high linguistic aptitude as well. This form included questions concerning (1) what the learners liked or did not like about the intervention, (2) how they found the structure of the worksheets, the language boxes, the historical sources, and the type of tasks, (3) whether they felt they learned

something (if yes, what in particular; if no, why), and (4) what we could improve in the future. As insights gained this way were helpful, it was decided to hand this feedback form to all those learners who were not interviewed at the end of cycle 3 (school A).

Moreover, to gauge potential development in terms of CDF use and history skills, students were asked to complete competency-based written tasks after each implementation phase. Here, rubrics were used to assess performance and track development. Finally, in a retrospective interview with the practitioner, we reflected on the process, revisited initial assumptions and conjectures in light of the results, and insights already gained throughout the process. Moreover, ideas for the adaption of the intervention and the fine-tuning of the materials were collected and discussed.

#### **5.3.4 Organisation of cycles**

As mentioned above, the basic structure of the cycles remained the same throughout the study, but each cycle pursued slightly different objectives, which entailed somewhat divergent realizations, set-ups, or evaluation foci.

The pilot cycles of this study (cycle 0) mainly aimed at field-testing the organizational sequence of the individual stages and first drafts of didactic materials as well as at piloting instruments. Additionally, this initial phase provided an opportunity to sharpen our focus, both concerning scope of research and materials. To be able to try out different approaches, two different groups (YA in school A, YB in school B) were involved in the pilot cycles. For example, in terms of social organisation, most of group YA's lessons consisted of group work and presentations, while YB's basic structure was teacher-guided talk as well as individual and pair work exercises. As the main focus of the pilot cycles was trying out the research design including its instruments and the first drafts of materials, exploration of the context was subsidiary. Therefore, the implementation phases were not videotaped at the pilot stage, also to reduce data, especially considering the low maturity level of the intervention. As a consequence, evaluation at this stage was mainly formative, focusing on internal structures, including soundness, feasibility, and local viability.

The sequence and organisation of the main cycles took into account the teachers' and students' availabilities while also ensuring purposeful development and fine-tuning of the intervention. The main objective of cycle 1 was to further develop some of the more general insights gained in the pilot phase as well as to expand ideas and attain new perspectives on how to improve and adapt our approach. At this stage, the focus of the evaluation process was still on formative aspects, centring on how working system components function in real contexts. Cycle 2, then, intended to examine how an already improved version of the same unit could work in a new context. Therefore, another group (B) at another school (B) was involved at this stage. In other words, in cycle 2, the intervention was further developed and applied in a new context to create more robust materials and pedagogical tools, also providing insights into institutionalisation, i.e., how an intervention is immersed in a wider educational organization such as a school (see McKenney & Reeves, 2014). Another shift in focus in this cycle was that summative evaluation now played a

more central role. Put differently, questions concerning possible ways of adaption and fine-tuning of interventions were increasingly replaced by questions about local effectiveness. Finally, in cycle 3, we built on the insights gained in all previous cycles to create a new unit (with a different topic), investigating whether the design principles developed thus far could be useful for creating another unit and whether a second intervention with the same group (A) would lead to more substantial learning gains compared to students who only experienced one unit (B). As cycle 3 concerned the same group as cycle 1, no needs analysis was required at this stage.

## 5.4 Data collection and instruments

In DBR literature, it is usually recommended to combine different data sources, data collection settings, and instruments to create more robust designs and to account for the complex and authentic settings typical for classroom research while also increasing the validity of the research project (see e.g. McKenney et al., 2006; McKenney & Reeves, 2012; Plomp, 2013). McKenney et al. (2006) further explain that triangulation of methods can address one of the issues frequently associated with DBR, namely the blurred roles of the researcher since they (co-)create materials they themselves then evaluate. By using multiple methods, possible conflicts of interest can be mitigated. Furthermore, the weaknesses of individual methods can be compensated to some degree (Knorr & Schramm, 2016; McKenney & Reeves, 2012). However, Knorr and Schramm (2016) point out that methods or data types should not only be accumulated but expediently integrated, which also ties in with McKenney et al.'s (2006) remark that the key point is not the amount of different methods or data types but the way they purposefully complement each other. These considerations also informed the design of the present study. Multiple methods were applied, and various data sources were included to make sure that different perspectives inform the answers to the research questions while not generating disproportionate amounts of data. An overview of the methods used in connection to the research questions is provided in Table 5 below.

Table 5. Research questions and methods

research questions (RQ)	method
RQ1: What kind of content-and-language-integrative pedagogical measures and materials (type and features) are needed to help students improve and elaborate their verbalization of cognitive processes (CDF use) as <ul style="list-style-type: none"> <li>(a) perceived by learners,</li> <li>(b) reported by teachers,</li> <li>(c) observed in written student performances?</li> </ul>	<ul style="list-style-type: none"> <li>– interviews (a, b)</li> <li>– design sessions (b)</li> <li>– written student performances (c)</li> </ul>
RQ2: How do students respond to explicit teaching of CDFs in the history CLIL classroom as <ul style="list-style-type: none"> <li>(a) reported by learners,</li> <li>(b) perceived by teachers?</li> </ul>	<ul style="list-style-type: none"> <li>– interviews (a, b)</li> <li>– written feedback (a)</li> <li>– (for triangulation: unstructured observations)</li> </ul>



RQ3: What is the effect of CDF-oriented teaching on the learners' development of historical competences and academic language skills as observed in written performances?	– written student performances (c)
---	------------------------------------

All these methods are also listed as typical and recommendable methods for DBR in McKenney and Reeves (2012).

In qualitative research in general (Caspari, 2016; Dörnyei, 2007) and DBR in particular (Euler, 2014; McKenney & Reeves, 2012, 2014), validity can be increased by leaving a detailed audit trail that is comprehensive and intersubjectively comprehensible. Therefore, thick descriptions of preconceptions, context, data collection, analysis, and reasoning were striven for in this dissertation. For that purpose, a research log accompanied the whole process. In connection to the importance of thick descriptions in DBR, Gravemeijer and Cobb (2006) argue for *virtual replicability* as a central quality criterion for DBR. They explain that DBR is basically a learning process of the whole research team, including practitioners, and thus a conventional conception of replicability does not make sense since a learning process is always a very subjective experience. However, if DBR researchers leave a detailed audit trail, they make it possible for others to track and comprehend the study and the insights gained from the data, enabling a replication of the learning process (Gravemeijer & Cobb, 2006). Freudenthal (1991) argues that “developmental research means experiencing the cyclic process of development and research so consciously, and reporting on it so candidly that it justifies itself, and the experience can be transmitted to others to become like their own experience” (p. 16). However, this does not mean that researchers replicating the learning process of others necessarily have to reach the same conclusions (Gravemeijer & Cobb, 2006).

Another way to increase validity in DBR is maximising ecological validity, which is a quality criterion somewhat defining DBR since researcher and teacher work closely together, making sure that theories and materials developed are viable for real-life use (McKenney et al., 2006). To this end, synergetic collaboration and clear communication between practitioner and researcher are vital to fully use both areas of expertise, as pointed out by Kelly (2006). Therefore, communication with participating teachers was continuously kept up in this study.

In the following sub-chapters, the individual data collection methods are introduced and described. All instruments can be found in [section I](#) of the appendix repository, whereas [section II/ D](#) provides an overview of the meta data of all data collection points.

#### 5.4.1 Interviews

Interviews were used to take account of the teachers' and students' perspective in terms of needs analysis, designing materials, and evaluation of interventions. According to Riemer (2016), interviews are a valuable method to access insiders' perspectives, including their opinions, experiences, convictions, and knowledge, but also traces of contradictory, vague, or unaware ideas and thoughts. Riemer (2016) explains that qualitative interviews also provide enough space for digging deeper via asking for elaboration, clarification, or explanation, resulting in rich and

possibly reflective and honest insights. Hence, conducting interviews seemed to be a useful method to include the teachers' and students' voices for the design of innovative teaching materials because meaningful design and re-design requires open and complex disclosure of participants' ideas and opinions. All interviews in this study were conducted in German, i.e., (one of) the participants' first language(s) or at least one of the two languages of instruction. The use of a familiar language was aimed at ensuring a relaxed atmosphere and enabling the interviewees to speak with no or only minimal language barrier.

In this study, each interview partner (or group of partners) was interviewed more than once, documenting their perspectives at different stages. Teacher A was interviewed four times, namely at the beginning and end of the pilot cycle and at the end of cycle 1 and 3. Teacher B1 was interviewed at the beginning and end of the pilot cycle, while teacher B2 was interviewed at the start and finish of cycle 2. All student groups were interviewed at the beginning and end of their respective cycle. Group A, the only group involved in two cycles (cycle 1 and 3), was interviewed three times: at the beginning of cycle 1 and the end of cycle 1 and 3. Especially in a project dedicated to development of some sort, which evidently seems to apply for material development, it made sense to conduct a series of interviews (see Dörnyei, 2007). Depending on the current stage, these interviews differed in set-up and purpose as is addressed in 5.4.1.1-4, where the four different interview types present in this study are outlined in the sequence as they occurred in the research cycles.

All interview types followed a semi-structured format, yet to differing degrees of structuredness due to their different purposes and set-ups. According to Dörnyei (2007), semi-structured interviews rely on an interview guide of pre-prepared (open) questions, which can be flexibly reordered and formulated. This, as Riemer (2016) further elaborates, allows for a more natural flow of conversation, which is more likely to encourage open talk and more in-depth answers. Dörnyei (2007) also explains that a semi-structured format is useful when the researcher already knows the field under consideration and therefore is able to prepare suitable questions beforehand and can flexibly "follow up interesting developments" (p. 202), which applies for the present study.

The interview guides (see [appendix section I/A](#)) were prepared drawing on methodological guidelines by Riemer (2016), Hermanns (2010), and Dörnyei (2007). For example, the interview guidelines start with easy, introductory questions to give the participants confidence and establish a positive atmosphere right at the beginning. Additionally, questions are kept short, comprehensible, and free of contradiction. Each question contains one idea and was formulated in an appropriate register, which would be semi-formal in the case of the present study, considering that these interviews were conducted with teenagers and former colleagues. The questions prepared were mainly open-ended and, if applicable, possible probes were prepared to go more into depth. At the end of the interview, the interviewees were given the opportunity to bring up anything they would like to address, and then the interviewer explicitly expressed gratitude for the interviewees' contributions and their time.

Dörnyei (2007) advises to pilot qualitative interviews. For that purpose, the interview guides were read by critical colleagues prior to the first pilot interviews to provide feedback especially in terms of overall length and length of individual questions, clarity of questions, possible redundancies, completeness of content, appropriateness of formulation, and whether there were any leading questions or contradictions. Comments and suggestions were then incorporated and after the first pilot interviews, participants were also asked for feedback concerning the questions and the way the interview was conducted. However, no further suggestions for improvement were made at this stage.

All interviews were recorded using Zoom H2N recording devices as well as a smartphone, and all participants were informed about this and consented to before starting the interview. A general introduction outlined the aim and scope of the research and the interview in particular. Strategies of anonymization were explained, and participants were reminded that there were no right or wrong answers. During the interview, recommendations for conducting interviews as described by Riemer (2016), Hermanns (2010), and Dörnyei (2007) were considered. As such, I tried to let the participant speak as freely as possible but guided them back to the topic in case they digressed too extensively. Also, using strategies of back-channelling and probe questions to encourage them to elaborate were used to ensure more in-depth answers. Furthermore, answers were occasionally rephrased to check whether I interpreted them appropriately. By doing these member checks, interpretative validity could be improved, as Dörnyei (2007) explains. Also, if a desirability bias was assumed, e.g., students saying something they thought their teacher would like to hear, mitigation thereof was aimed for.

#### **5.4.1.1 Semi-structured interviews with teachers: needs analysis**

In the needs analysis phase, interviews with individual teachers were conducted to establish their point of departure, including some personal background information, views on CLIL (history) teaching and on the role of language for content teaching, as well as current practices. Moreover, in these semi-structured interviews, first ideas for material design were collected. These areas of inquiry were already set prior to the interview; therefore, this interview type was rather on the structured, focused half of the spectrum. Yet some degree of flexibility was desirable to create a productive environment and to encourage comprehensive and extensive input.

#### **5.4.1.2 Focus group interviews with students: needs analysis**

As this study does not only consider teachers' perspectives but also students' views, semi-structured interviews were conducted with students for the needs analyses as well. However, students were not interviewed individually but in groups for two reasons: First of all, Dörnyei (2007) as well as Riemer (2016) explain that group interviews can create synergistic environments that are enjoyable and stimulating, yielding rich and insightful data. Cohen, Manion, and Morrison (2011) add that especially when working with children or teenagers, a group set-up could help reduce shyness and intimidation. Moreover, a group setting could also decrease reluctance because the participants can collectively argue but also challenge each other (Cohen et

al., 2011). In other words, focus group interviews do not only provide individuals' perspectives but also the interaction within a group. Secondly, interviewing students in groups was time-efficient, especially given that the individual perspectives of students were not essential per se, unlike with teachers, who played a more project-defining role.

Similar to the interviews with teachers, an interview guide following a semi-structured format was prepared, centring on essential areas of interest, such as views on CLIL, CLIL materials, reflections on prior experiences, the role of language for content subjects, and learning needs in terms of content and language. To account for the group setting, however, the guide for the students is slightly more open than the teachers' to facilitate a natural flow of discussion and interaction. During preparation, the same principles as for the teacher interviews were considered. Additionally, principles for focus group interviews as specified by Bohnsack (2010) were taken into account. Bohnsack's (2010) principles coincide with most of Dörnyei's (2007) recommendations. However, Bohnsack (2010) suggests not to assign turns and always address the whole group, while Dörnyei (2007) advises to give space to each participant and avoid dominance of one participant. Particularly with teenagers, it seemed to make more sense to also address the quiet ones because they would otherwise not volunteer. What is more, the interventions should not only be tailored to extrovert students, which is especially important in the context of FL use. To ensure a productive and respectful discussion, the introduction before the interview included rules of conduct (e.g., we listen to each other, we respect each other's views, etc.).

Choice of focus students was made on two grounds: First, within-case sampling tried to strive for maximum variation in terms of performance in history and English. For that purpose, the results of the pre-intervention written performances were considered. Second, students should be willing to share their opinions and be able to reflect on learning and teaching to ensure insightful reflection sessions. Based on the preliminary results of the written tasks, I created a list of potential interview partners and discussed it with the teacher, who then chose participants best fulfilling our criteria (i.e., maximum variation in terms of competence, willingness to share and reflect). Nonetheless, voluntariness, of course, was crucial, not only for ethical reasons but also, as mentioned above, because willingness to share views is critical for meaningful outcomes.

#### **5.4.1.3 Retrospective focus group interview with students: evaluation**

The retrospective interview is a method used to collect the participants' views, insights, and reflections after a specific action (Riemer, 2016), which is the implementation of the intervention in this case. The retrospective focus group interview with students followed the same basic premises as the group interviews of the needs analysis with one difference: The retrospective interview was less focused with only a number of pre-prepared guiding questions. Instead, there was a greater focus on open reflection on the intervention including concrete materials. Riemer (2016), for example, suggests using medial support to facilitate recall. To this end, printouts of all materials were given to the students along with green and red pens. Students were asked to mark everything they liked in green and everything they did not particularly like in red. They were

encouraged to discuss their choices with their peers and note down comments why they chose the respective colour.

Retrospective interviews also provided opportunities to conduct member checks, which increase interpretative validity, as Dörnyei (2007) or McKenney et al. (2006) explain. As such, retrospective interviews could be used to bring up ideas from previous data collections and discuss them once again, checking whether the researcher interpreted something correctly and/or whether the participants' views had changed since we last spoke. For this purpose, the interview guides were enriched with notes prior to the interview.

#### **5.4.1.4 Retrospective interviews with teachers: evaluation**

The retrospective interview with the teacher marked the last step of each data collection cycle. Again, some prepared questions were formulated to ensure that no major areas were neglected. These main areas were the teacher's overall perspective on the intervention, observations concerning students and their learning processes, material evaluation, ideas for revision and further interventions, and process evaluation including aspects of teamwork. This last area was important to ensure future positive collaboration as stressed by McKenney et al. (2006), who argue that a successful DBR project needs to be mutually beneficial. Despite having some questions prepared in advance, the focus of these interviews was definitely on the reflection of the teacher and, to some extent, mine too, considering that we were both involved in the creation of the intervention. In preparation of the retrospective interviews with teachers, all pertinent data of the respective cycle was browsed and everything that seemed relevant was added in keywords to the interview guide. Moreover, the printouts of the materials with the students' comments and notes in red and green were brought to the teacher's interview, asking them for their opinions and how they experienced working with these materials. As such, this type of interview was the least structured one of the four types present in this study since the process was very open and dependent on the flow of the conversation and reflection.

#### **5.4.2 Competency-based written tasks**

Competency-based written tasks were employed to track the students' performances at different stages. In the early phases of the project, the task's purpose was mostly diagnostic, while at later stages, the students' results also provided insights into local effectiveness. To be able to follow trail of progress and compare results, the tasks adhered to the same structure but were not too repetitive in terms of content (topic and type of visual source, if possible). Additionally, the tasks used in this study were based on the format of the Austrian final history exam to ensure that Austrian learners were familiar with the format and because it could be assumed that teachers, learners, and their parents would perceive such practice as purposeful. An advantage of using competency-based tasks was that the structure of the tasks could remain the same while topics could be changed. It was presumed that using this general structure more than once would not lead to considerable memory effects detrimental for this study. In a way, procedural knowledge how to approach these types of tasks was even desirable, as it helped learners develop their

historical competences and prepare for their final exams at the same time. In this sense, the degree of maturation was of interest in this dissertation.

The tasks used for this study were constructed on the basis of guidelines for in-service teachers issued by the Austrian Ministry of Education (2011, 2013) and Ammerer and Windischbauer (2011). These publications outline and discuss the structure of competency-based history exams and their content in the form of historical and political competences. Moreover, these publications provide a list of performative verbs and a comprehensive collection of sample tasks. Following these guidelines, tasks should always include a visual, textual, or audio (-visual) historical source which is not only illustrative but the central element of the task. Tasks can be either concept/topic-oriented or methods/genre-oriented. Since the same task structures should be used up to three times in the present study, it made sense to detach the tasks from the notional, topic-oriented level and rather centre them on procedural knowledge which can be applied to any given topic. Thus, the methods/genre-oriented type was chosen. Moreover, for the sake of comparability, only one type of source was considered. To avoid differing levels of linguistic input as well as technical challenges, textual and audio (-visual) sources were excluded, leaving visual sources as the source type to be used in this study.

As recommended by the guidelines mentioned above, the tasks were designed to reflect the competences defined in the Austrian history curriculum. In this study, two competences were chosen as main focal points, namely deconstruction competence as well as orientation competence. A previous study by Bauer-Marschallinger (2016) could show that these two competences play a central role in the Austrian history curriculum and competency-based history testing since both of them are inherently connected to historical source analysis, which is the main precept of the final exam.<sup>26</sup> What is more, in other competency-models of other countries (see, for instance, NCHS, 1996; Union of German History Teachers, 2006), these skill types are central too.

As already mentioned in chapter 4, the Austrian guidelines prescribe that all tasks should contain task items on three different levels (Kühberger, 2011; Ammerer & Windischbauer, 2013):

- I. *Reproduction*: On the lowest level, demonstrating declarative knowledge, such as naming facts and defining terms, and reproductive use of methods, e.g., identifying sources, or discriminating between source types, are central.
- II. *Reorganisation/ transfer*: At level II, students autonomously work with sources, such as when explaining relations, organizing content (*reorganisation*), or applying appropriate methodological scripts (*transfer*).
- III. *Reflection/problem-solving*: Level III demands reflexive and self-reliant approaches. This means that students at this level are able to critically reflect on historical insights, connections between different aspects, as well as their own methodological choices and outcomes (*reflection*). Moreover, students should be able to justify their evaluations, interpretations, and reasoning (*problem-solving*).

---

<sup>26</sup> See subchapter 4.2 for more information on these competences and section 4.3.3 concerning the final history exam.

Kühberger (2011) specified a set of *Operatoren* for each of these levels, i.e., performative verbs which are intended to prompt pre-defined and practised actions (see section 4.2.5 for more information).

Based on these considerations, a task template was created (see [appendix section I/B/ file 2](#) and [3](#)), consisting of seven tasks for one visual source in the pilot phase. All three levels are present in this template, ranging from reproduction (I), to transfer (II) or reorganisation (II), to reflection (III) or problem-solving (III). For each item, the level, its central operation, and the target competences were specified. Moreover, target CDF types were defined on two levels: *CDF-episode* refers to the overall main function, while *CDF-basic* indicates possible sub-components (see subchapter 3.4 for more information). A sample item is given below:

Item no. 7 (later no. 5):

[reflection (III): discussion – relevance]  
Argue whether (or in which ways) this [source type]/ [issue depicted] is still relevant in the 21<sup>st</sup> century. (*EVALUATE-episode*, *REPORT-basic*, *CATEGORIZE-basic*, *DESCRIBE-basic*, *orientation*)

This item targets orientation competence as it asks the learners to connect past and present and evaluate to which extent or in what ways an artefact of the past might still be relevant for our present existence. When making such judgements (*EVALUATE*), historians often *REPORT* past and present circumstances and/or developments and establish differences and similarities of the two timelines (*CATEGORIZE*). Ideally, historians would not only compare historical context of the source and present developments, but they would also directly relate the historical context to what is depicted in the source, for which one would *DESCRIBE* a source. In other words, a historian might also implicitly gauge the validity of a source (i.e., the extent the contents shown correspond to the historical context) when evaluating its historical relevance. These task templates can then be filled in relation to the source given. For example, for the topic of the Industrial Revolution (pilot cycle), a drawing depicting a 19<sup>th</sup>-century factory (see Arnould, 1876) was chosen. The relevant item then looks like the following:

Item no. 7 (5)/ pilot:

[reflection (III): discussion – relevance]  
Argue whether (or in which ways) this picture might still be relevant in the 21<sup>st</sup> century. (*EVALUATE-episode*, *REPORT-basic*, *CATEGORIZE-basic*, *DESCRIBE-basic*, *orientation*)

In the pilot study, these tasks were field-tested in two groups at two points in time. The topics covered were the American Revolution, the Industrial Revolution, and ideologies of the 19<sup>th</sup> century (see [appendix section I/B/ file 4](#)). Item no. 2, which is concerned with the historical contextualisation of the visual source, was frequently left out or misunderstood in the pilot phase. Therefore, this item was excluded from the main study. Item no. 6, which asks students to *DEFINE* a concept relevant for the topic of the source, was also eliminated because it does not fit to the overall focus on deconstruction and orientation competence. During pilot analysis, the list of target CDF-basic types was also revised and expanded, since sometimes the learners produced

CDF compositions that I did not envisage but nonetheless made sense from a subject-specific perspective. Finally, it should be mentioned that the initial template, all concrete tasks, and changes proposed were discussed with the teachers, validating the template and individual tasks from an ecological perspective.

### 5.4.3 Intervention design-sessions

All intervention design-sessions were audiotaped using a Zoom H2N recorder and a smartphone. In these sessions, the teacher and I collaboratively prepared, discussed, evaluated, and improved the plans and materials of the intervention. The exact nature of these design sessions depended on the overall stage of the project (see 7.1.2 for a more detailed description of the design sessions and the workflow). In line with principles of DBR, we voiced our thoughts explicitly, negotiating the interests of practice and research while never losing sight of the overall aim of the project (see Euler, 2014). Moreover, we discussed assumed user competences required for the implementation of the intervention, potential critical events, and their solutions, as well as the possibility of adjustments under changeable conditions as recommended by Euler (2014). Apart from the final versions of the lesson plans and materials, drafts and revisions can be found in the appendix repository, [section III \(didactic materials\)](#), documenting the development of the designs.

All these meetings happened face-to-face, and communication between personal meetings was kept up via e-mail and telephone. Moreover, as already mentioned in subsection 5.3.3.2, the teachers and I shared all material via a file-share repository, making visible any changes in real-time.

### 5.4.4 Observations (online/ offline)

In this study, observations were used for two different purposes, namely gathering information about the context and documenting the implementation. As for the purpose of establishing a comprehensive picture of the context, Cohen et al. (2011) state that observing allows researchers to collect in-situ data in authentic situations to inform on various levels of context, such as the physical, social, interactional, or programme setting. Since thick descriptions and comprehensive contextual knowledge are crucial for theorizing in DBR, it seems that observations are a useful tool to collect and complement information on the contextual level. What is more, unlike interviews, observations do not rely on self-reported or second-hand information but facilitate more direct data collection (Dörnyei, 2007). This entails that observations permit noticing events that might otherwise be left unsaid by participants for whatever reason, as Cohen et al. (2011) point out.

At this point it should be stressed that this method was not used as a main research tool but rather as a means to ensure rich contextual information, especially in terms of what other methods could not address directly. Therefore, an unstructured approach was favoured over a structured approach so that insights could be captured that were not anticipated or expected beforehand or were maybe even consciously or unconsciously withheld by the participants.



Schramm and Schwab (2016) report that research literature often distinguishes between *participating* and *non-participating* observation and refer to Johnson and Christensen (2012), who identified four types on a spectrum, namely *complete participant*, *participant-as-observer*, *observer-as-participant*, and *complete observer*. In this study, the role of *observer-as-participant* was chosen, as I did not intend to fully immerse myself in the field and only planned for a limited amount of time present in the classroom. Moreover, students and teachers were aware of and consented to the observation, unlike with a *complete-observer* approach.

Another frequent distinction in research literature, according to Schramm and Schwab (2016), is *online/ offline*, referring to the possibility of either observing in-situ or later using audio or video recordings. In the case of the pilot study and the needs analysis, notes were made in-situ, i.e., online, and expanded afterwards, while ongoing reflections and tentative records of continuous interpretations accompanied the whole process as suggested by Cohen et al. (2011) in order to leave a detailed audit trail. As for documenting the implementation during the main research cycles, both offline and online approaches were employed. Offline observations enabled the documentation and traceability of the classroom activities. For the purpose of traceability, a video camera was placed in the classroom to record the implementation phase and sequence of events. Moreover, audio recording devices (Zoom H2N) were placed next to the teacher and selected students. The audio recording device for the students was moved whenever appropriate during the lessons to capture the contributions of different students. Additional online observation was aimed at collecting ‘live’ information on the context and impressions from an observer’s point of view similar to the observations during the needs analysis.

## 5.5 Data analysis

Since various data collection methods were employed in this study, several methods of analysis were needed too, as outlined in Table 6 below.

Table 6. Research questions and methods of data collection and analysis

research questions (RQ)	method of collection	method of analysis
RQ1: What kind of content-and-language-integrative pedagogical measures and materials (type and features) are needed to help students improve and elaborate their verbalization of cognitive processes (CDF use) as (a) perceived by learners, (b) reported by teachers, (c) observed in written student performances?	interviews (a, b)	qualitative content analysis (QCA)
	design sessions (b)	
	written student performances (c)	CDF-based coding and linguistic rating subject-specific rating
RQ2: How do students respond to explicit teaching of CDFs in the history CLIL classroom as (a) reported by learners, (b) perceived by teachers?	interviews (a, b) written feedback (a)	qualitative content analysis (QCA)
	(for triangulation: unstructured observations)	partial analysis of field notes and lesson transcripts (critical episodes)

RQ3: What is the effect of CDF-oriented teaching on the learners' development of historical competences and academic language skills as observed in written performances?	written student performances (c)	CDF-based coding and linguistic rating subject-specific rating
---	----------------------------------	---

Just like data collection, data analysis in DBR rests on transparent and tight audit trails of analysis. Using and explicating the use of conceptual frameworks for analysis and interpretation leads to increased interpretative validity according to Gravemeijer and Cobb (2006). An explicit and traceable process of analysis and interpretation provides others with the opportunity to generalize to other cases (case-to-case generalization) or to broader theory (analytic generalization), which in turn increases external validity as McKenney et al. (2006) argue. Another strategy to increase validity in qualitative analysis mentioned by Dörnyei (2007) is to examine outliers and to consider alternative explanations in order to counteract potential researcher bias. All these considerations were kept in mind when analysing the data collected in this study.

In what follows, transcription conventions are outlined and then each method of analysis is described and discussed in detail.

### 5.5.1 Transcription

Oral data was transcribed based on the guidelines outlined in Kuckartz (2016), who expanded and complemented the guidelines by Dresing and Pehl (2015). These guidelines are focused on content and do not consider a number of verbal and non-verbal features in order to ensure a feasible process as well as a readable and comprehensible transcript. This loss of linguistic-analytical depth is acceptable because these aspects were not part of the research focus.

The transcription conventions based on Kuckartz (2016) and Dresing and Pehl (2015) have been specified and expanded in the course of transcribing pilot data. The list of conventions and rules can be found in the appendix, [section I/ F](#). In the pilot study, I did the transcription myself but since the schedule was tight and results of previous cycles affected later cycles, transcription in the main cycles was done by a qualified external person. By doing the first round of transcribing myself, I could provide the second transcriber with enough examples of how I envisioned the transcription. Moreover, I remained in contact with the second transcriber, answering any arising questions and helping out in cases of uncertainty (e.g., special terms unknown to the transcriber, contextual information needed, etc.).

### 5.5.2 Qualitative content analysis

Qualitative content analysis (QCA) was used in this study to analyse interview data, design-session transcripts, and written student feedback. QCA allows systematic, yet to some degree flexible analysis of large amounts of communicative data and its reduction in relation to the focus of research (Schreier, 2012). Therefore, QCA seemed suitable to process and analyse the communicative data gained in this study. While there are several other well-known approaches to QCA (e.g., Mayring, 2015, or Schreier, 2012), I opted for Kuckartz's (2016) approach since he

not only presents a well-structured and feasible procedure but also developed a software, MaxQDA (VERBI Software, 2017, 2019), thus ensuring high compatibility between the theoretical construct and software tools.

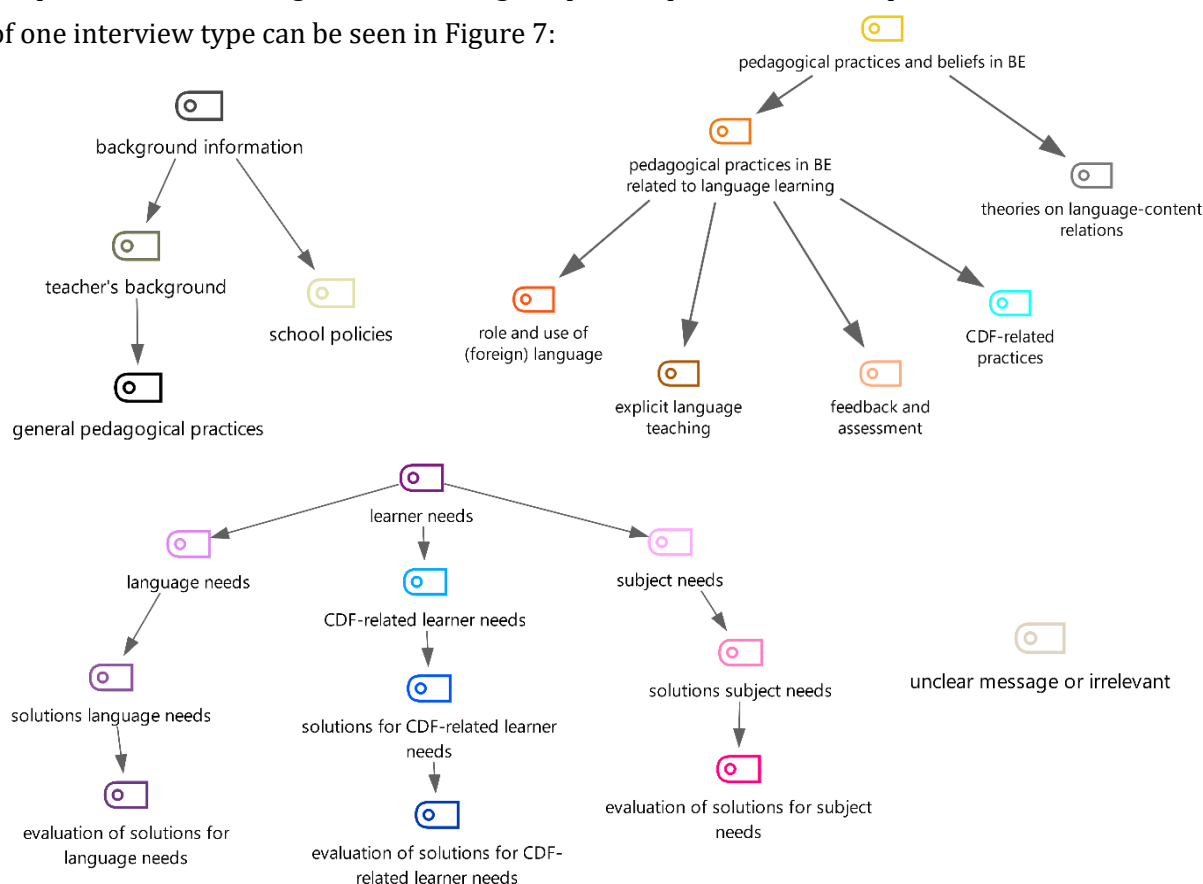
Looking at the type of QCA, this study mostly made use of structuring QCA, which, Kuckartz (2016) explains, is well suited for identifying, describing, and relating main topics and sub-themes in semi-structured interviews, focus groups, and other communicative data too, and as such, structuring content analysis works well with thematic and theoretical categories (as mostly used in this study). Burwitz-Melzer and Steininger (2016) further state that structuring QCA permits both case-oriented as well as theme-oriented analyses via the construction of thematic matrices, thus facilitating comprehensive, yet focused analysis. Depending on the interview type conducted for this study (see section 5.4.1), the focus was either on both case-orientation and theme-orientation or only on theme-orientation. For example, the analysis of teacher interviews both considered aspects specific for this case and for particular topics, while the analysis of interviews with students was only focused on topics and not on individual cases. As mentioned already, teachers played a very productive role in the research design; therefore, their individual cases were of interest. Individual students, on the other hand, did not actively participate in the concrete design process, and they also provided their input collectively. Consequently, student data was not examined from a case-focused point of view.

In the retrospective interviews, which aimed at evaluating the intervention, the codebook does not only contain thematic or theoretical categories like for the interviews of the needs analysis or for the design sessions but also very basic evaluative codes, namely *positive evaluation*, *negative evaluation*, and *inconclusive evaluation*. These codes were added when the interviewees (or respondents of the written feedback survey) expressed appraisal of any of the aspects (covered by thematic or theoretical codes) connected with the intervention, e.g., *complexity of tasks* or *educational value of tasks*. This way, it was possible to systematize and visualize which aspects were mostly perceived as positive or negative by creating code co-occurrence models as well as crosstabs.

A core aspect of QCA is the formation of categories. Berelson (1952) even goes as far as stating that “a content analysis can be no better than its system of categories” (p. 147). Therefore, forming a coding scheme of categories and sub-categories needs careful consideration and should always be based on the focus of research (Kuckartz, 2016). Research literature usually distinguishes between deductive and inductive categorisation as well as hybrid forms (e.g. Burwitz-Melzer & Steininger, 2016; Kuckartz, 2016). Kuckartz (2016) reports that deductive-inductive hybrid forms are the most common type for the development of coding schemes. In the case of combining deductive and inductive categories, one usually starts out with deductive, a-priori categories, directly deduced from the interview guide and/or theory, and then one would expand and differentiate the coding scheme based on the material in several steps (Kuckartz, 2016). This was also the general strategy in this study, mainly adhering to the phases and guidelines for structuring QCA as outlined by Kuckartz (2016):

- (1) initiating textual work, highlighting passages, and writing up memos;
- (2) defining main categories directly based on the interview guide and elaborating and differentiating them based on 10-25% of the material;
- (3) first round of coding (main categories);
- (4) compiling all passages coded for the same category;
- (5) inductive defining of sub-categories;
- (6) second round of coding (main categories and sub-categories);
- (7) simple and complex analysis and visualisation.

The coding systems and coding guides were developed on the basis of the pilot data (step 1-6). The data of the main study was then analysed with these tools, meaning that only step 6 and 7 as outlined above were conducted in the main study. As for the balance of deductive and inductive categories, the more structured and therefore more focused interviews of the needs analysis relied more on a-priori categories based on the interview guide, while the design sessions and retrospective interviews of the evaluation phase were much more open, thus requiring more inductive categories. Kuckartz (2016) explains that in bigger studies or studies that have already progressed somewhat, it is possible to already define some sub-categories deductively too, but these need to be checked against the data and, if necessary, adapted based on the empirical sources. This was also the case in this study, especially for the semi-structured interviews. Here, some sub-categories were already established a-priori since the interview guides were quite detailed. Some additional sub-categories were found inductively, and the majority of the pre-established sub-categories were revised or specified on the basis of pilot data. Moreover, the system of these categories was re-arranged based on empirical sources. In other words, based on the pilot data, some categories were merged, split, or specified. An example of the final code trees of one interview type can be seen in Figure 7:



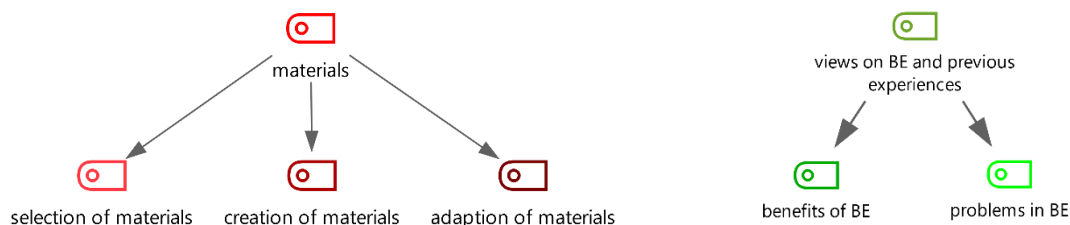


Figure 7. Code trees: needs analysis teacher interview

All other code systems can be found in the appendix, [section I/ C \(codebooks and code trees\)](#).

Each code was defined, illustrated with appropriate examples taken from the data, and then systematized into codebooks following Kuckartz's (2016) guidelines. These codebooks for all interview types and design sessions can be found in the appendix repository, [section I/ C](#). Here is a sample code-definition taken from the needs analysis teacher interview codebook:

The code "theories on language-content relations" refers to any theoretical considerations and reflections on how language and content learning/ teaching are related or independent; e.g., in what ways subject-specific learning goals are dependent on language skills, etc. This code also covers any theorizing and reflecting on the concept of CDFs and performative verbs. Moreover, it also includes reflections on how these theories have changed throughout the teacher's career and what has caused these changes but not how this shows in their teaching practice. Here, appropriate practice-focused codes should be used.

Contributions relevant for this code may be prompted by the researcher. Answers reacting to any thought-provoking impulses are also included here.

Example:

R: [...] wie denkst du, oder denkst du, dass sprachliche und fachliche Bedürfnisse zusammenhängen? [How do you think, or do you think that language and content needs relate?]

TA: Ja, aufgrund der Zentralmatura beziehungsweise der Operatoren auf alle Fälle (.) mittlerweile. [Yes, because of centralized final testing and the performative verbs, for sure (.) by now.]

As the example shows, these code definitions describe in detail when a certain code applies. If applicable, it includes which aspects can be expected, which questions might prompt answers relevant for this code, whether reactions to the researcher's input are included, or in which cases the code would not apply.

Kuckartz (2016) points out that some text passages can be coded with several categories, unless, of course, the codes are constructed to rule each other out. Schreier (2012) does not agree as she argues that categories need to be mutually exclusive. However, especially with thematic categories, such an approach is not expedient since some passages, even individual sentences, can contain more than one (sub-)topic (Kuckartz, 2016). At the same time, it seems crucial for a comprehensible analysis to segment according to units of meaning, which excludes the possibility to separate one unit just to facilitate two codings in one unit of meaning (Kuckartz, 2016). Thus, in this study, text passages could be coded with multiple codes if deemed appropriate to avoid splitting units of meaning while not losing analytical substance. What is more, in the case of the retrospective interviews, the evaluative categories (*positive, negative, inconclusive evaluation*) were conceptualized as add-on codes. Schreier (2013) further argues that all text passages need

to be coded but adds that for passages irrelevant to the focus of the study, one could create “residual categories” (p. 175). To cover the interview data in its entirety, the present study, too, added a category for the participants’ utterances that were *irrelevant*, such as small talk or side talk, or *unclear*, i.e., when the meaning of a passage could not be decoded (see codebooks in the [appendix section I/ C](#)).

After coding in MaxQDA 2018 (VERBI Software, 2017) and later MaxQDA2020 (VERBI Software, 2019), thematic matrices were constructed and exported to spreadsheets (Microsoft Excel). There, summaries of all topics were written to facilitate comparison of compact data within and between topics and cases. These summary tables can be found in the appendix repository, section II/ [A \(interviews\)](#) & [C \(design sessions\)](#). Moreover, the results of the analyses in MaxQDA were visualized in the form of code co-occurrence models, hierarchical codes/sub-codes models with frequencies, and code matrix browsers, visualizing frequencies and overlapping of codes. Written student feedback was considered as an additional case in the summary tables and the code matrix browsers. In the code co-occurrence models and hierarchical codes/sub-codes models, however, the data of interviews and written feedback were combined.

### 5.5.3 Coding and rating of written tasks

To be able to make claims about the quality of the students’ written performances in terms of historical competences, as targeted by RQ3, two assessment grids were developed since, unfortunately, no official standardized assessment grids for competency-based history testing have been published so far by the Austrian educational authorities. Therefore, the proposed rubrics should be considered as a working tool that only serves to operationalize formative and summative evaluation in the context of this specific research project. As such, their practicality for pedagogical use has not been factored in during the development process. Nonetheless, I would like to highlight that this would represent an important research gap since assessment in CLIL has been and still is a major concern of CLIL teachers (del Pozo & Llinares, 2021; Morton, 2020; Otto & Estrada, 2019; see also section 7.1.1).

In this study, the written student performances were analysed with the help of a CDF-based coding and rating scheme and a subject-specific rating scheme (see [appendix section I/D - rubrics](#)). The two rubrics were developed in the following way: Initial drafts of domains, levels, and descriptors were derived from theory. The linguistic rubric is based on the CDF construct and central notions for historical literacy (see chapter 3), whereas the history-focused rating scheme is based on the FUER model and the testing guidelines for history testing (see chapter 4). These initial rubrics were tested on pilot data and subsequently further refined. Apart from specifications for the different descriptors, it was decided that zero levels were necessary too. Moreover, an initial coding and rating guide was drafted to clarify the procedure. For instance, first coding and rating CDF use followed by rating the subject-specific domains was found to be the most efficient sequence and was thus fixed in the manual. Throughout coding and rating the written performances of the main cycles, further specifications to both the rubrics and the coding and

rating guide turned out to be necessary. In other words, the rubrics and the manual were only finished after rating all performances of the main cycles; thus, all performances had to be rated once again. More information about the different rubrics and a description of the respective rating processes are presented below in subsection 5.5.3.1 and 5.5.3.2, respectively. Then, in subsection 5.5.3.3, statistical procedures are outlined. This subsection also includes information on the intrarating procedures aimed at determining the reliability of the ratings. While reliability checks have been performed, I would like to point out that the rubrics have not been benchmarked or systematically validated, as this would have exceeded the limits of the present dissertation. Here, the reader is referred to the PhD project by del Pozo (in progress) focused on integrated assessment in CLIL history education, working with the CDF construct as well (see also del Pozo & Llinares, 2021). In general, more research into integrated ways of assessment would be warranted, including integrated assessment that promotes learning opportunities (deBoer & Leontjev, 2020; Morton, 2020)

### 5.5.3.1 The CDF-based coding and linguistic rating

The linguistic rubric (Figure 8, next page) consists of three dimensions, namely *use of CDF types*, *use of linking devices*, and *use of linguistic markers* typical of historical discourse (*hedging & nominalisation*). These dimensions each comprise two sub-dimensions, for which three levels plus a zero level were defined, ranging from no/hardly to very good control in the respective aspect. This way, initial, mid-stage, and final written task performances could be compared. The number of levels was developed on the basis of the pilot data and with a view to the subject-specific rubrics, which also consist of three stages plus a zero level, reflecting the guidelines for history testing (as outlined in section 5.4.2).

Starting with CDF use, CDF realizations were coded on two levels (episode & basic) based on the identification of underlying communicative intentions as defined in Dalton-Puffer's (2013, 2016).<sup>27</sup> The dimension *use of CDF types* is differentiated into *choice* and *composition of CDF types*. *Choice of CDF types* describes to what extent the CDFs used correspond to the target episodes and basic CDF types as specified in the task template. *Composition of CDF types* is concerned with whether or to what degree the assembling of individual basic CDFs is logical and comprehensible and to what extent these compositions sustain the CDF episode, i.e., the overarching communicative intention.

As for *appropriateness of linking*, the rating scheme differentiates between appropriateness in terms of *form* and *function*. *Form* ratings pay attention to orthography, collocations, and syntax, whereas punctuation is not taken into account. *Function* ratings are concerned with whether the linking strategies used correspond to the CDF employed, i.e., the function of the episode. Both sub-dimensions do not only consider appropriateness but degree of general linkage, i.e., how often learners link their ideas explicitly (e.g., "ideas are well linked, and linking is linguistically accurate"/ level 3).

---

<sup>27</sup> See subchapter 3.4 for more information on Dalton-Puffer's (2013, 2016) construct.

		level 0	level 1	level 2	level 3	individual levels	levels of sub-categories	overall level
use of CDF types	choice of CDF types	no/hardly any CDF types are target CDF types	some CDF types are target CDF types	most CDF types are target CDF types	all episodes and most basic CDF types are target CDF types			
	composition of CDF types	composition of CDF types is illogical/ unclear	composition of CDF types is partly logical/ clear	composition of CDF types is generally logical/ clear	composition of CDF types is logical/ clear throughout			
use of linking devices	appropriateness of linking in terms of <u>function</u>	no/hardly any appropriate linking (in relation to CDF type)	some appropriate linking (in relation to CDF type)	ideas are generally linked appropriately (in relation to CDF type)	ideas are linked appropriately (in relation to CDF type)			
	appropriateness of linking in terms of <u>form</u>	no linking/ linking not linguistically accurate	some linking is linguistically accurate	ideas are linked, and linking is generally linguistically accurate	ideas are well linked, and linking is linguistically accurate			
use of linguistic markers typical for historical discourse (hedging/ nominalisation)	appropriate use of hedging	no/ hardly any evidence of hedging/ inappropriate use of hedging	some evidence of hedging but may be partly used inappropriately	hedging is present and mostly used appropriately	hedging is clearly present and used appropriately			
	appropriate use of nominalisation	no/ hardly any evidence of nominalisation/ inappropriate use of nominalisation	some evidence of nominalisation but may be partly used inappropriately	nominalisation is present and mostly used appropriately	nominalisation is clearly present and used appropriately			

Figure 8. Linguistic rubric

Starting with CDF use, CDF realizations were coded on two levels (episode & basic) based on the identification of underlying communicative intentions as defined in Dalton-Puffer's (2013, 2016).<sup>28</sup> The dimension *use of CDF types* is differentiated into *choice* and *composition of CDF types*. *Choice of CDF types* describes to what extent the CDFs used correspond to the target episodes and basic CDF types as specified in the task template. *Composition of CDF types* is concerned with whether or to what degree the assembling of individual basic CDFs is logical and comprehensible and to what extent these compositions sustain the CDF episode, i.e., the overarching communicative intention.

As for *appropriateness of linking*, the rating scheme differentiates between appropriateness in terms of *form* and *function*. *Form* ratings pay attention to orthography, collocations, and syntax, whereas punctuation is not taken into account. *Function* ratings are concerned with whether the linking strategies used correspond to the CDF employed, i.e., the function of the episode. Both sub-dimensions do not only consider appropriateness but degree of general linkage, i.e., how often learners link their ideas explicitly (e.g., "ideas are well linked, and linking is linguistically accurate"/ level 3).

<sup>28</sup> See subchapter 3.4 for more information on Dalton-Puffer's (2013, 2016) construct.



Turning to features of historical discourse, *nominalisation* and *hedging* were chosen as features in focus due to their importance in the field of history and their relatively compact nature (see section 3.3.2). Again, the descriptors for these two sub-dimensions consider frequency and appropriateness (e.g., “hedging is clearly present and used appropriately”/ level 2). As specified in the coding and rating manual, hedging devices include the occurrence of modal verbs, conditionals, probabilistic adverbs like “probably”, “possibly” when used to qualify statements, and qualifying verbs such as “seem”, “assume”, “appear”. As for nominalisations, all types were considered, i.e., words that contain a derivational affix or zero derivation functioning as a noun, including gerunds. However, nominalisations taken from the prompt were not counted.

The coding and rating procedure is specified in detail in the coding and rating manual ([appendix section I/ B/ file 5 - coding support](#)) and involves the following steps: first, in MaxQDA, CDF realizations were coded on basic and episode level, and attempts at and uses of nominalisations, hedging, and linking were highlighted in specific colours. These results were documented for each task in the CDF note sheet (see Figure 9) using abbreviations defined in the rating manual.

task	CDF episode	CDF basic						organisation	linking (function)	linking (form)	hedging	nominalisation
1	DS											
2	EA		EO									
3	EV		RE		DS		EA					
4	EV		RE		DS							
5	EV		RE		CA		DS					

Figure 9. Notes sheet for the linguistic rating

To make the process manageable in terms of workload and overview, the different items were not rated individually but holistically in the rubric. This means that the results in the notes sheet were combined and matched to the descriptions of the CDF rubric. Thoughts and decisions were further tracked and illustrated by examples in the general research log and memos in MaxQDA. The rating manual further provided information concerning specific items, ambiguous cases, and concrete examples and was consulted in case of doubt. The ratings for the different students were collected and documented in Excel spreadsheets, where also group averages were calculated. All numerical data was then imported into SPSS for further examination (see subsection 5.5.3.3).

### 5.5.3.2 Subject-specific rating

The subject-specific rubrics used for this study are based on the guidelines for history testing, which in turn are based on the works by the FUER group (Ammerer & Windischbauer, 2011; Austrian Federal Ministry for Education, 2011; Körber et al., 2007). These guidelines have already been used for the creation of the task template, thereby ensuring coherence of tasks and assessment tools. While the template specified the target competences and the levels of historical thinking skills, the rubrics focus on whether students actually performed the competence targeted by the task and whether they did so on the appropriate level as defined in the template. These criteria were then transformed into three rubrics, one focusing on level of thinking skills

(reproduction – reorganisation/transfer – problem-solving/reflection, see section 4.2.5), termed *level rating*, and one each for the two competences included in the prompt (deconstruction competence and orientation competence, see section 4.2.1), labelled as *competence rating*. These three rubrics are split into three different dimensions, which all, in turn, define three stages plus a zero stage. The dimensions of the thinking-level focused rubric are summarized below and presented in Figure 10:

- *Target level* indicates to what extent the student's answer reflects the intended thinking level, i.e., whether an appropriate historical thinking skill was performed fully, partly, to some extent, or not at all. For example, did learners really analyse (thinking level II) or did they only reproduce declarative knowledge (thinking level I)?
- *Accuracy/ relevance* scores determine to which degree the reported facts (reproduction, level I), reorganisations and transfers (level II), or reflections and solutions (level III) are relevant to the task and historically correct or justified (i.e., whether or to which degree they contradict established historical developments).
- *Systematicity*, which is also an important factor in the FUEr grading logic and testing guidelines, describes to what degree these reports, analyses, reflections, etc., are systematic (or not).

	stage 3	stage 2	stage 1	stage 0
target level	The answer <b><u>reflects or exceeds the intended level</u></b> ; i.e., the right action is performed; level 1 = reproduction, level 2 = transfer/reorganisation, level 3 = reflection/ problem-solving	The answer <b><u>reflects the intended level for the most part</u></b> ; i.e., the right action is performed but <b><u>not exclusively/ fully</u></b> ; level 1 = reproduction, level 2 = transfer/reorganisation, level 3 = reflection/ problem-solving	The answer partly <b><u>reflects the intended level to some extent</u></b> ; i.e., the right action is performed <b><u>some of the time or to some extent</u></b> ; in case of level 3 tasks, level 2 is performed most of the time	The answer <b><u>does not reflect the intended level</u></b> ; i.e., the right action is <b><u>not performed</u></b> ; level 1 = reproduction, level 2 = transfer/reorganisation, level 3 = reflection/ problem-solving
accuracy/ relevance	on level I: the reported facts are <b><u>correct/ relevant</u></b> for the task  on level II: the reorganisations, transfers, etc., are <b><u>historically accurate and relevant</u></b> for the task  on level III: reflections and evaluations are <b><u>justified</u></b> (i.e., do not contradict established historical developments) and <b><u>relevant</u></b> for the task	on level I: the reported facts are <b><u>mostly correct and/ or relevant</u></b> for the task  on level II: the reorganisations, transfers, etc., are <b><u>mostly historically accurate and/ or relevant</u></b> for the task  on level III: reflections and evaluations are <b><u>mostly justified</u></b> (i.e., rarely contradict established historical developments) <b><u>and/ or relevant</u></b> for the task	on level I: the reported facts are <b><u>partly incorrect and/ or irrelevant</u></b> for the task  on level II: the reorganisations, transfers, etc., are <b><u>partly historically inaccurate and/ or irrelevant</u></b> for the task  on level III: reflections and evaluations are <b><u>partly unjustified</u></b> (i.e., sometimes contradict established historical developments) <b><u>and/ or irrelevant</u></b> for the task	on level I: the reported facts are <b><u>incorrect and/ or irrelevant</u></b> for the task  on level II: the reorganisations, transfers, etc., are <b><u>historically inaccurate and/ or irrelevant</u></b> for the task  on level III: reflections and evaluations are <b><u>unjustified</u></b> (i.e., contradict established historical developments) <b><u>and/ or irrelevant</u></b> for the task
systematicity	reports/ reorganisations/ reflections, etc., are <b><u>systematic</u></b>	reports/ reorganisations/ reflections, etc., are <b><u>mostly systematic</u></b>	reports/ reorganisations/ reflections, etc., are <b><u>mostly unsystematic</u></b>	reports/ reorganisations/ reflections, etc., are <b><u>unsystematic</u></b>

Figure 10. Level rating rubric

The second area of assessment concerns the demonstration of the target competence. Again, three dimensions are specified on three stages plus a zero stage for both competences, namely:

- *Target competence* determines to what extent an answer demonstrates control of the target competence. For deconstruction tasks, this includes to what extent the source is deconstructed in terms of the aspect in question and how one refers to the source (directly, indirectly, generically, incorrectly, not at all). Similarly, orientation-oriented descriptors measure how deeply and explicitly one engages with the contents of the source in terms of discussing historical relevance of the concepts displayed in the source.
- *Justification/ comprehensibility* measures the degree of justification (fully justified, with some effort, great effort, not justified) and whether the connection between source and answer (deconstruction competence) or source and present (orientation competence) is tangible and comprehensible. In other words, are answers fully justified or are their justifications vague, implied, or non-existent and are their answers comprehensible or reasonable?
- *Scope* scores indicate the amount of detail (great amount, sufficient amount, some details, no details) and whether all parts are covered or not.

Both rubrics are presented below (Figure 11 & Figure 12). To avoid confusion, it should be noted that the terms “level I, II, III” here reflect the guidelines for history testing (= level of thinking skill), while *stages 0 to 3* are the levels specified for this research project. To ensure some degree of comparability to the levels defined by guidelines, a three-stage hierarchy was constructed here as well.

	stage 3	stage 2	stage 1	stage 0
target competence	the source is <b>deconstructed</b> in terms of the <b>aspect in question</b> ; including <b>direct reference</b> to the <b>source</b> and its <b>content</b>	the source is <b>deconstructed mostly</b> in terms of the <b>aspect in question</b> , either with <b>indirect reference</b> to the concepts displayed (e.g., by talking about the relevant aspect) or <b>generic reference</b> (e.g., "as the picture shows")	deconstruction of the aspect in question is not visible, but <b>related historical concepts are discussed</b> ; either with <b>no reference</b> to the source or <b>(incorrect) generic reference</b> to the source (e.g., "as the picture shows")	<b>no deconstruction</b> of the source, including <b>no discussion of related historical concepts</b> and <b>without reference</b> to the source
justification/ comprehensibility	the connection between source and answer is <b>fully</b> comprehensible/ justified	the connection between source and answer is <b>tangible/ justified with some effort</b> (i.e., the answer is <b>very comprehensible/ reasonable</b> , but the justification is <b>vague or superficial</b> )	the connection between source and answer is <b>comprehensible/ justified with great effort</b> (i.e., the answer is <b>somewhat comprehensible/ reasonable</b> but there is <b>no (explicit) justification</b> )	there is <b>no (justified) connection</b> between source and answer/ <b>no deconstruction of the source</b>
scope	<b>great</b> amount of detail, <b>all</b> parts covered	<b>most</b> parts covered, <b>sufficient</b> amount of details	<b>some</b> parts covered, supported by <b>some</b> details; OR a <b>good</b> amount of details, but <b>central point is not included</b>	<b>no</b> details, <b>substantial</b> parts missing

Figure 11. Competence rubric targeting deconstruction competence

	stage 3	stage 2	stage 1	stage 0
target competence	there is a <u>direct</u> connection between the contents of the source and the present/ historical relevance of the source is <u>explicitly</u> discussed	the <u>connection between concepts displayed in the source and the present</u> are <u>sufficiently</u> discussed/ historical relevance of these concepts is <u>either discussed without direct reference</u> to the source <u>or with generic reference</u> (e.g., "as shown in the picture")	connections between <u>concepts related to the source and the present</u> are <u>discussed in terms of historical relevance</u> ; either with <u>no reference</u> to the source or <u>(incorrect) generic reference</u> (e.g., "as shown in the picture")	there is <u>no</u> connection between source and today/ <u>no</u> discussion of historical relevance
justification/ comprehensibility	the <u>connection between source and the present</u> is <u>fully</u> comprehensible/ justified	the <u>connection between source and the present</u> is <u>tangible/ justified with some effort</u> (i.e., the answer is <u>very</u> comprehensible/ reasonable, but the <u>justification is vague or superficial</u> )	the <u>connection between source and the present</u> is <u>comprehensible/ justified with great effort</u> (i.e., the answer is <u>somewhat</u> comprehensible/ reasonable, but there is <u>no (explicit) justification</u> )	there is <u>no (justified)</u> connection between source and the present/ no discussion of historical relevance
scope	<u>great</u> amount of detail, <u>all</u> parts covered	<u>most</u> parts covered, <u>sufficient</u> amount of details	<u>some</u> parts covered, supported by <u>some</u> details; OR a <u>good</u> amount of details, but <u>central point is not included</u>	<u>no</u> details, <u>substantial</u> parts missing

Figure 12. Competence rubric focusing on orientation competence

Keeping in mind that students were confronted with items on different levels and competences, it was decided that global assessment is not expedient. Instead, assessment was separated for the individual items, which also ensures transparency. These ratings were collected and documented in Excel spreadsheets, where averages of the individual descriptors (i.e., the results of one descriptor for all five task items) were calculated automatically. Figure 13 presents the spreadsheet that was used to document the ratings.

			LEVEL				COMPETENCE			
task	level of task	target comp.	target level	accuracy/ relevance of content	systematicity	overall level stage per item	target comp.	justification/ comprehensibility	scope of content	overall competence stage per item
1	I	DC								
2	II	DC								
3	II	DC/OC								
4	III	DC								
5	III	OC								
average per descriptor										
			overall level stage				overall competence stage			

Figure 13. Spreadsheet to document history ratings

Moreover, averages were determined for the individual students and the whole group. These results were then fed into SPSS for further examination. Like with the linguistic rating, reflections and thoughts were tracked in the research log as well as with the help of MaxQDA memos.

### 5.5.3.3 Statistical procedures

As mentioned above, results of all ratings were documented and calculated in Microsoft Excel spreadsheets. The results for the individual descriptors were automatically transposed into average content and average language outcomes. Moreover, the results of the 12 different descriptors were summarized in tables listing all students and calculating means for the different descriptors for each data set. Although the concept of stages/ levels would imply a hierarchal, ordinal system, thus requiring working with integers and medians, it was decided to work with means for the following reasons: Even though the rating scale is technically ordinal, it is common practice to treat such scales as metric because the distances between the individual ratings are theoretically the same and there is a zero level as well (Cleve & Lämmel, 2016; Field, 2017). For instance, school grades, or any rating requiring some degree of subjective judgement, are also technically ordinal, yet many practitioner and researchers treat them as metric data by assuming that the distances between the grades or levels are equal, consequently allowing them to calculate averages, *t*-tests, ratios, etc. (Albert & Marx, 2016). Similarly, I treat my scales as metric variables in order to be able to calculate mean values and *t*-tests and also to consider extreme cases (which medians do not reflect).

All results were imported into SPSS 26.0 (IBM Corp., 2019). The following measures were calculated for all data sets and are reported in the appendix, [section II \(data analysis\)/ B \(pre- and post-intervention tasks\)](#):

**Standard descriptive values** (e.g., mode, median (*Mdn*), mean (*M*), standard deviation (*SD*), variance, range (*R*), minimum (*Min*), maximum (*Max*), etc.) were calculated for overall results (content, language, overall), all descriptors individually (CDF-related and history-related), and word counts for the purpose of providing detailed descriptions of the results and a frame for contextualization.

**Tests of normality** were performed for overall results (content, language, overall), all individual descriptors (CDF-related and history-related), and word counts as a basis for further statistical decisions.

**Plots** were produced to visualize results and confirm or reject normal distribution. Histograms and boxplots were created for all overall results and all individual descriptors, whereas stem-and-leaf plots were only produced for linguistic descriptors as this type of visualization works better with integers (or at least a manageable number of different results).

**Extreme values** of overall results (content, language, overall) were reported to contextualize other statistical values and provide further details. For all other descriptors, extreme values were not reported as these lists turned out to be multitudinous without providing information that could not also be deduced from plots.

**Correlation coefficients** were calculated and reported for all descriptors and average overall language and content results for the purpose of investigating potential relations between the various rating dimensions. For normally distributed data, Pearson's *r* was used, whereas

correlations of non-normal data were examined via calculating Kendall's tau b ( $\tau_b$ ). Like Spearman's rho ( $\rho$ ), which is often used for this purpose, Kendall's correlation coefficient can be used for non-parametric data; yet Kendall's tau is assumed to have stronger statistical properties than Spearman's rho especially with small samples (Field, 2017). Any statistically significant correlations were marked in shades of green in the correlation analysis tables (the darker the shade, the stronger the correlation). Correlation tests between the overall category (e.g., *overall content*) and their various descriptors (e.g., *target level*) are presented in italics and in pale colours because their results would be distorted given that these descriptors were used for calculating overall outcomes (i.e., a certain degree of correlation could be expected from the outset). Nevertheless, *overall content* and *overall language* results were included in this table to investigate to which extent individual descriptors might correlate with overall results of the other domain (e.g., *overall content* results and *choice of CDF type*).

**Tests of comparison** were performed to measure whether differences between the pre- and post-tasks are statistically significant, i.e., not a product of chance (most likely). To calculate differences and their significance between pre- and post-task, i.e., the results at two points in time, paired samples *t*-tests were conducted for normally distributed data and Wilcoxon signed-rank tests for non-normally distributed data. These are standard tests of comparison for pre- and post-test comparisons with consistent group compositions (Field, 2017). For all statistical tests, an alpha level of .05 was used. Additionally, the effect size was calculated using Cohen's *d* for *t*-tests and Wilcoxon effect size  $r_w$ , as suggested by Field (2017). In group A, students completed tasks at three points in time. Therefore, different tests were needed. For normally distributed data, the standard choice is ANOVA with repeated measures (Field, 2017), which was conducted for all descriptors where all three data sets presented normal distributions. Additionally, *Mauchly's Test of Sphericity* was employed to check whether corrections were needed (Field, 2017). In all ANOVA calculations, sphericity was not violated; therefore, no corrections to the degrees of freedom (*df*) were necessary. For one-way ANOVA with repeated measures, three different types of effect sizes have been suggested, namely partial eta squared ( $\eta^2_p$ ), generalized eta squared ( $\eta^2_g$ ) (Lakens, 2013; Olejnik & Algina, 2003), and omega squared (Field, 2017). Partial eta squared can be calculated in SPSS and is therefore widely used; however, it has been found to be misleading and imprecise (Field, 2017; Lakens, 2013; Olejnik & Algina, 2003). Thus, Field (2017) recommends omega squared, which Olejnik and Algina (2003), however, describe as misleading too. They, together with Lakens (2013), recommend generalized eta squared instead, which can be calculated with the help of the open access software The Jamovi Project (2021). Consequently, effect sizes for ANOVA results were reported in the form generalized eta squared ( $\eta^2_g$ ) but, in the appendix, also via partial eta squared ( $\eta^2_p$ ) to reflect common practices in the discipline. For non-normally distributed data, Friedman tests were conducted to measure the significance of differences (Field, 2017). To analyse effect sizes of Friedman tests, Field (2017) recommends conducting a series of Wilcoxon signed-rank tests, reporting  $r_w$  for each comparison.

To measure the reliability of the ratings, 20 written student performances were randomly selected (i.e., four texts per data set, accumulating to 22.5% of all texts collected) and rated again after eight to 14 months since the original rating. Unfortunately, no other rater with appropriate expertise in both history didactics (the FUER model in particular) and linguistics (CDFs especially) was available. Thus, I opted for an intrarater approach (see Bortz & Döring, 2016; Knorr & Schramm, 2016; Kuckartz, 2016). **Intrarater reliability** was determined in the following way: Pearson's correlation (bivariate correlation) between the results of the original and the second rating were calculated for parametric data (see Bortz & Döring, 2016). These generally show a high correlation, ranging from  $r_p = .70$  ( $p = .001$ ) for *systematicity* (history) to  $r_p = .97$  ( $p < .001$ ) for *overall history*, indicating a good match between first and repeated rating. However, Person's  $r$  does not consider (consistent) shifts, meaning that Pearson's  $r$  can be perfect even if the second rating would always be higher/lower than the first as long as it does so to the same degree (Bortz & Döring, 2016). Therefore, for normally distributed data (assumed as metric data), *intraclass correlation (ICC)*, which is sensitive to absolute agreement, was calculated too. Here, correlations range from  $ICC = .79$  ( $p < .001$ ) for *systematicity* to  $ICC = .98$  ( $p < .001$ ) for *overall content*. According to Bortz and Döring (2016), ICCs of 0.7 and above can be considered as indication of good reliability. For non-normally distributed data, Kendall's coefficients were calculated. Results range from  $\tau = .65$  ( $p = .001$ ) for *linking in terms of form* to  $\tau = .77$  ( $p < .001$ ) for *nominalisation*. Interestingly, history ratings, generally, were more reliable than linguistic ratings. Nonetheless, overall ratings, (i.e., *overall content* results and *overall language* results) both indicate very high reliability ( $ICC = .98$ ,  $p < .001$ , for content and  $ICC = .97$ ,  $p < .001$ , for language). The reason for this might be that individual descriptors derived from the same concept were not distinctive enough and therefore might have influenced each other. Yet, overall, they seem to balance each other out, leading to very convincing and reliable overall results. More information on the intrarater reliability analysis can be found in the appendix repository, [section II \(data analysis\)/ B \(pre- & post-tasks\)/ file 5 \(intrarater reliability analysis\)](#).

#### 5.5.4 Field notes and lesson transcripts

As already argued in section 5.4.2, in-situ observations were conducted to gather information about the context and also during the implementation phase. Here, the focus was on situational information that other methods could not capture. Therefore, observations were unstructured, collecting anything potentially noteworthy. After the lessons, these field notes were solidified and complemented with post-observation reflections. Later in the project, these field notes, along with the video recordings, were used to substantiate the transcripts of the lessons. In other words, the transcripts were annotated with the help of the videos and field notes, adding anything relevant to the implementation of the intervention that was not audible on the audio tracks. Finally, critical episodes, i.e., sequences relevant to understand the interviewees' reflections and evaluations as voiced in the interviews, were identified in the expanded transcripts and selected for discussion, corroborating the participants' perspective and observed behaviour.

## 6. Pilot study

In the summer term of 2018, I conducted two pilot cycles (= cycle 0), one in each participating school, in order to field-test the organisational sequence of the individual stages and the research instruments for data collection, improving the research design for the main cycles (see subchapter 5.4). For example, after the piloting, the written task template was adapted and shortened, removing two items that turned out to be problematic or redundant, respectively. Interview guidelines used for different types of semi-structured interviews were only marginally revised (e.g., wording, typos). Moreover, the data collected in the pilot cycles enabled the development and improvement of the analytical instruments for this project (see subchapter 5.5). On the basis of the written performances of the pilot groups, theory-based rubrics were defined more precisely, ensuring that the descriptors were relevant for the data at hand and specific enough to apply. For instance, this process has shown that a zero-level was necessary for all domains. Additionally, the interview data and the design session recordings allowed to clarify the transcription conventions so that another person could transcribe the main data set reliably and consistently. The transcripts of the pilot cycles (produced by the researcher) could be used to develop the coding systems for the qualitative content analysis for each interview type and the design sessions of the main study.

In addition, the pilot phase provided an opportunity to try out the first designs of the intervention, sharpening focus and type of materials by trying out different formats and types of tasks with different groups (group YA with T<sub>A</sub> in school A, group YB with T<sub>B1</sub> in school B). The topic of the pilot didactic unit was ideologies of the 19<sup>th</sup> century (conservatism, liberalism, Marxism/socialism, and capitalism) and was planned for four to six lessons (see [appendix section III/ C - pilot units](#)). These first drafts were designed by the researcher and shown to colleagues for feedback. Moreover, during a professional workshop focused on CLIL, the drafts were presented to international researchers and local teachers, collecting feedback as well. Their comments were considered in a first revision before discussing the materials with T<sub>A</sub>.

The process of the first pilot cycle in school A is outlined in Bauer-Marschallinger (2019) with a focus on research methodology and teaching materials, i.e., to what extent DBR can put theory into practice in the context of CLIL and what CDF-focused CLIL materials could look like. In this publication, some initial findings were presented too. Yet, it needs to be stressed that the pilot cycles were a work in progress, and thus the results should be treated with caution. First insights of this pilot cycle were then considered for the second draft of the materials for the second pilot cycle in school B. One of these insights, for example, concerned the social organisation of the didactic unit. In school A, we divided up the four ideologies so that different learner groups would research and elaborate on tasks concerning only one of the four ideologies each. This process was scaffolded in small steps, using explicit linguistic support often based on specific CDFs to help learners understand sources and express their thoughts. At a later stage, the different groups presented 'their' topic, including an exemplary source analysis, to their peers. Here, we found that the groups learned a lot about their own topic but struggled to understand what their peers



presented. Considering the newness of the approach and the abstractness and complexity of the content, it was decided to abandon the idea of these ‘expert groups’ and have everybody work on the same content with frequent breaks between learner-centred episodes, allowing the teacher to promptly clarify issues and to ensure that everybody stays on track. Moreover, it was decided that these learner-centred activities should be a mix of individual, pair, and group work in the following cycle. The revised materials were then implemented by T<sub>B1</sub> with group YB in school B. Unlike the first pilot cycle, the process and results of the pilot cycle in school B have not been published. Taking into account the nature and purpose of the pilot data and the volume of the data of the main study, presenting these results in detail does not seem appropriate. Instead, I would like to report central insights gained in both pilot cycles that informed the upcoming main cycles and which might be relevant when discussing the results of the main study.

First of all, both student groups reported that they were struggling with expressing complex historical content precisely but could not really put their finger on what it was that they needed exactly except for a larger lexical base. In both groups, they mentioned that when faced with performative verbs in tests, they would just write whatever comes to their mind since it was not quite clear what they were supposed to do precisely, some even welcoming the “leeway” (“Spielraum”) these performative verbs would supposedly provide. Their teachers, too, reported that learners often failed to react adequately to performative verbs. In general, the students did not seem to be very much aware of how language and content might be connected and did not appear to fully grasp the communicative intentions of (some) performative verbs used in tests. Teachers, both being language teachers too, were quite convinced of the role of language for teaching content subjects but, thus far, had not paid much attention to linguistic aspects in the context of CLIL.

After the intervention, students and teachers of both groups seemed to accept a CDF-based approach to consider language-related aspects in content lessons, appreciating the general approach and most of the tasks. Learners and teachers highlighted the learner-centred nature of the tasks and scaffolding, helping the learners engage with the sources and promoting understanding of complex tasks. However, some learners reported that they felt slightly overwhelmed, especially in school A, where the unit was largely organized in ‘expert groups’. In this context, they would have appreciated more teacher guidance, considering that they were not used to focusing on language in a content subject. To allow for more teacher guidance while ensuring in-depth treatment of tasks as well as avoiding overload, it was decided with the teachers to sacrifice some breadth of content.

Furthermore, the learners perceived certain tasks as redundant, which were mostly those tasks that focused on language quite explicitly. Here, the students missed the relation to content learning. Some language boxes, too, were perceived to be too detached from content, overcomplicating things from their perspective. Therefore, the tasks would need to combine language and content more genuinely and, at the same time, communicate educational purposes more explicitly in a learner-appropriate way. In some cases, the tasks were also not entirely clear,

or their execution was not practical. What was felt to be missing was a proper, and in the case of group YB, a long enough closing activity to ensure and promote uptake. All these comments were noted and considered in following design sessions.

When observing classes in both contexts, it also became apparent that the teachers were not used to teaching linguistic and subject-literacy-related aspects explicitly. Therefore, it seemed crucial, especially in the beginning phase of such a project, that researcher and teacher communicate more and that the materials include more didactic information for the teacher (e.g., through a teacher's version with explicit commentary) to ensure that the teacher knows what to consider when employing these materials for the first time.

In terms of results of the written tasks, group YA improved moderately in terms of content (average gain 18%, from  $M = 1.55$  at T1 to 1.84 at T2) and slightly in terms of language (average gain 10%, from  $M = 1.52$  at T1 to 1.67 at T2).<sup>29</sup> Group YB decreased slightly in both aspects (-8% for content, from  $M = 2.05$  at T1 to 1.89 at T2, and -4% for language, from  $M = 1.68$  at T1 to 1.62 at T2). Here one contextual variable that needs to be considered is that the post-intervention task was conducted in the last lesson of the day and directly after a German exam, also coinciding with the last day to finalize assessment.<sup>30</sup> As a consequence, they might not have been as focused as they would usually be. Here, one lesson for upcoming data collections was to plan more wisely and communicate more with school partners to be better informed about contextual variables. Unfortunately, in the pilot cycles, there was simply too little time left in the semester for such considerations. In any case, none of these results are statistically significant as measured by Wilcoxon signed-rank tests.<sup>31</sup> This is not surprising considering that the number of participants completing both pre- and post-test was very low (10 out of 12 in group YA and 12 out of 19 in group YB). In terms of effect size, however, a medium effect can be reported for the changes in content rating of both groups, i.e., the increase in group YA and the decrease in group YB; see footnote 31 below. Concerning changes in language ratings, a small effect of  $r_w = .27$  was observed in group YB. Yet it needs to be kept in mind that based on these cycles, the rubrics and the task template were revised and improved. For these reasons, the results of the pilot tests should be treated with extreme caution, and therefore going into detail would be futile. The purpose of these first written tasks was to field-test and improve the template as well as the rating frameworks and not to test the efficiency of the intervention. Typically for DBR, these steps would follow later (see, for instance, McKenney & Reeves, 2012, or Euler, 2014).

---

<sup>29</sup> The results of items that were later discarded are not considered in these numbers. Moreover, only those students that completed pre- and post-intervention tasks are considered here.

<sup>30</sup> This group finished the school year earlier due to a major school trip (work placement) in June.

<sup>31</sup> Wilcoxon signed-rank test results for YA:

- content:  $T = 28.0$ ,  $p = .161$ ,  $r_w = .31$
- language:  $T = 21.0$ ,  $p = .236$ ,  $r_w = .27$

Wilcoxon signed-rank test results for YB:

- content:  $T = 19.0$ ,  $p = .117$ ,  $r_w = .32$
- language:  $T = 23.0$ ,  $p = .645$ ,  $r_w = .09$

T-tests were not conducted due to the small sample size.

## 7. Analysis

In the following chapter, the findings of the empirical study will be presented. Reflecting the process of design studies, this chapter is structured according to the four main phases of a research cycle, namely *needs analysis* (7.1), *designing the intervention* (7.2), *implementation phase* (7.3), and *evaluation of interventions* (7.4).

### 7.1 Needs analysis

This subchapter presents the results of the needs analysis of both groups involved in this study, thus providing an empirical basis for answering RQ1:

What kind of content-and-language-integrative pedagogical measures and materials (type and features) are needed to help students improve and elaborate their verbalization of cognitive processes (CDF use) as

- (d) perceived by learners,
- (e) reported by teachers,
- (f) observed in written student performances?

Furthermore, such a needs analysis ensures an in-depth contextualization of the study, which is a crucial component of any qualitative research but especially in DBR due to its demand for creating ecologically valid designs and for enabling case-to-case generalization (McKenney & Reeves, 2012; Van den Akker et al., 2006a).

First, participants' views are systematically presented and qualitatively analysed based on the interviews with teachers (7.1.1) and students (7.1.2). Subsequently, section 7.1.3 examines the results of the pre-intervention written tasks both from a history-didactic and linguistic perspective. Every section of the needs analysis concludes with a summary of results relevant for designing the intervention. In these summary subsections, central links to other studies contextualize the findings. This, however, is not a full discussion, which will follow in chapter 8 of this thesis, synthesizing the findings of all research phases and the literature reviewed.

To guarantee a transparent and comprehensible analysis, analytical steps and products are available in [section II of the digital appendix \(data analysis\)](#). As such, this section of the appendix contains summary tables, MaxQDA code-matrices, hierarchical codes/sub-codes models, and code co-occurrence models of all interviews ([subsection A](#)). As for the pre-intervention tasks, SPSS calculations (descriptive statistics, tests of normality, correlation tests, and *t*-tests) can be viewed ([subsection B](#)). Rating spreadsheets (Excel files) as well as original data and their respective qualitative analyses (MaxQDA files) can be accessed via personal request to ensure that the data is only used for authorized research purposes.

### 7.1.1 Semi-structured interviews with the teachers

The initial interviews with the teachers were aimed at taking account of the teachers' perspectives on and experiences with BE<sup>32</sup> as well as at identifying their needs and, indirectly, those of their students. Since T<sub>A</sub> participated in the pilot study, this interview was already conducted in March 2018, i.e., the semester preceding the main study. The interview with T<sub>B2</sub> took place in March 2019. This section now outlines and compares main themes and aspects relevant for the design of the interventions, which were identified in the QCA of the two initial interviews with the teachers. These results are grouped in four subsections, namely *pedagogical practices and beliefs related to language learning* (7.1.1.1), *materials* (7.1.1.2), and *learner needs* (7.1.1.3), followed by *summary and implications for the design* (7.1.1.4).

#### 7.1.1.1 Pedagogical practices and beliefs related to language

To better contextualize the teachers' practices, it is crucial to understand their view of language and the role of English in their bilingual history classroom. In this respect, the teachers gave rather different answers. T<sub>A</sub>, despite being an English teacher, said that language was simply a medium and not a learning objective per se:

1	English translation <sup>33</sup>	Original quote
T <sub>A</sub>	<b>Simply for conveying content, nothing else.</b>	<i>Einfach nur zur Übermittlung des Inhaltes, keine andere.</i>

This might be connected to the policies of her programme and how she was socialised there. When asked whether one could separate content and language learning, rather than elaborating on her thoughts, she reported school policies of the past and how it had affected her practice:

2	English translation	Original quote
T <sub>A</sub>	Initially, when I started teaching, it <u>was</u> like that [= separating content and language]. And <b>this was really a learning process for me because, being an English teacher, I was used to immediately<sup>34</sup> correcting pronunciation or tenses, for example</b> (laughs). But I stopped doing that and now I don't know where we are headed to (laughs).	<i>Das war am Anfang, wie zu unterrichten begonnen habe, <u>war</u> das so. Und das war für mich ein Lernprozess, weil ich automatisch als Englischlehrer zum Beispiel falsche pronunciation oder falsche Zeit habe ich <u>sofort</u> verbessert (lacht). Da bin ich aber jetzt weggegangen, und jetzt weiß ich nicht mehr [...] wohin der Weg geht (lacht).</i>

Later, when I inquired into her language-related practices more specifically, she again explained that she was indoctrinated not to be a language teacher in content subjects when she started teaching bilingually:

<sup>32</sup> In school A, the educational programme is labelled "bilingual", whereas in school B, the label CLIL is used. In the interviews, I used the label the participants were familiar with. To summarize both these programmes, the abbreviation BE (bilingual education) is used in this part of the thesis.

<sup>33</sup> All translations were done by the author. These translations follow the wording of the original, but, for readability, exclude hesitation markers and other features of spoken language that do not add any meaning (as judged by the author).

<sup>34</sup> Underlined passages were stressed vocally in the interview (or lesson), whereas **bold type** indicates the author's emphasis. All (other) transcription conventions can be reviewed in the appendix ([section I/ F](#)).

3	English translation	Original quote
T <sub>A</sub>	During my first years, I was inculcated not to teach language in content subjects. <b>I was always told ‘content teaching is not language teaching’</b> , and at the beginning, this was really <u>hard</u> for me, but I got used to it. <b>And now we’ll see</b> (laughs).	<i>Das ist mir so eingetrichtert worden in den ersten Jahren. Hat immer geheißen sprach- ah nein der Fachunterricht ist kein Sprachunterricht und das ist wirklich <u>schwer</u> am Anfang. Aber wie gesagt, mittlerweile habe ich mich daran gewöhnt und jetzt schauen wir mal (lacht).</i>

Based on this understanding, T<sub>A</sub> said that she would neither assess language nor teach language explicitly. Yet, she mentioned two times that she did not know how her practices as well as school policies would develop in the future. This indicates that T<sub>A</sub> views current policies as unsatisfactory and might welcome clearer guidelines that reflect the current state of knowledge.

Despite her rejection of explicit language instruction in content subjects, she said that she taught field-specific and general academic vocabulary, both in English and German, due to the learners’ perceived gaps in vocabulary. On the whole, her approach to FL learning in the bilingual classroom seemed implicit save for vocabulary teaching. In this vein, T<sub>A</sub> said that she ensured a great amount of exposure to the FL and that she encouraged her students to speak English during group work:

4	English translation	Original quote
T <sub>A</sub>	Well, <b>I certainly teach more in English</b> than [German], I definitely don’t do 50/50.	<i>Also ich unterrichte sicher mehr Englisch als [Deutsch], ich mache sicher nicht 50/50.</i>
	TA: <b>I keep reminding them to use English as their working language</b> , also when they talk to each other. R: Okay, so, you go around and and TA: = Yes, yes, exactly. Basically, just like in my language classes.	<i>TA: Also darauf weise ich sie schon hin, dass sie Englisch als Arbeitssprache verwenden sollen, auch wenn sie miteinander diskutieren. R: Okay also du gehst durch und und TA: =Ja, ja, genau. Im Endeffekt wie im Sprachunterricht mache ich es dann.</i>

T<sub>B2</sub>, on the other hand, reported that he used English not as frequently but for specific topics, aspects, or learning phases. This interviewee stated that he only taught in English if it brought added value, e.g., through the use of authentic source material or if it made sense for the topic at hand more generally:

5	English translation	Original quote
T <sub>B2</sub>	that’s the great thing about CLIL materials, there’s <b>added value</b> for them [= the learners] because we can use original sources. [...] <b>If I can’t see the added value</b> per se at all, and we just do something in English, <b>then I struggle in terms of motivation.</b>	<i>das ist das Tolle mit den CLIL Materialien, dass es für sie ein Mehrwert ist, weil wir ja Originalquellen dann verwenden können [...] Wenn ich den Mehrwert per se überhaupt nicht seh und einfach naja gut, dann machen wir halt etwas auf Englisch, dann tu ich mir auch motivationstechnisch ein bisschen schwerer</i>
	I think <b>English has to ‘lend itself’ somehow.</b>	<i>ich finde, das Englische muss sich auch irgendwie anbieten.</i>

He stressed a number of times that English sources and materials really enriched his teaching, but he would not want to teach in English for the sake of the FL. What is not entirely clear from this extract is whether the motivation issues mentioned here relate to himself, his students, or everybody involved. Considering the overall tone and content of the interview, it can be safely

assumed that he referred to his students, but the way he put it could imply that he was included too.

Another important criterion for his choice of language was the complexity of the content. He argued:

6	English translation	Original quote
T <sub>B2</sub>	<b>when reaching a certain level of difficulty, I tend to switch to German to explain connections.</b> This is also what many colleagues say, summaries and introductions in English, that's great, but once it's really about the topics where students struggle with the content immensely, then <b>a second language would be another obstacle.</b>	<i>ab einem gewissen Schwierigkeitslevel merke ich auch, dass ich dann tendenziell ins Deutsche überschwenk, um die Zusammenhänge zu erklären. Das ist auch das, was viele Kollegen und Kolleginnen oft sagen, so Zusammenfassungen, Einstiege in Englisch sehr gut, aber wenn's dann um Themen geht, wo die Schülerinnen und Schüler wirklich auch massive inhaltliche Schwächen haben, dass dann die zweite Sprache hier doch noch einmal ein Stolperstein sein kann.</i>

When dealing with complex content, this teacher suggested just doing one little aspect or only the revision in English, while the larger part of the unit should remain in German. Also, he explained that when they ran out of time, he usually shortened the English parts.

As for the language used by the students, T<sub>B2</sub> stated that he accepted both English and German for answers within teacher-student-talk as well as during group or pair work. He said that he would not reject German even if the whole unit was currently in English but would try to elicit or model English answers. He stressed that he would rather have the students talk freely than having them associate negative feeling with the language or the topic, as that would impede learning:

7	English translation	Original quote
T <sub>B2</sub>	consequently, they always have the possibility in my classes [...] 'I can't do it in English, I'll say it in German'; that won't be the end of the world. Sometimes, I try to paraphrase [...] or [I ask] 'does anybody have a clue how to say it in English?'. But nobody should be alienated [...] Honestly, I think this would have a deterring effect [...] <b>And if they start associating something negative</b> [with the language], <b>the learning effect will be zero anyways.</b>	<i>dementsprechend haben sie bei mir immer auch die Möglichkeit [...] das kann ich jetzt nicht auf Englisch, ich sag's auf Deutsch, geht, geht die Welt nicht unter. Manchmal versuch ich's zu paraphrasieren [...] oder hat jemand eine Ahnung, wie könnte man das noch gut auf Englisch sagen? Aber es wird kein Schüler von mir jetzt irgendwie vorn Kopf gestoßen [...] ich seh da auch ehrlich gesagt eher eine abschreckende Wirkung [...] Und wenn sie anfangen damit was Negatives zu verbinden, ist der Lerneffekt jetzt sowieso gleich Null.</i>

Turning now to his evaluation of the role of language learning for his subjects, his views seemed to somewhat differ from T<sub>A</sub>'s opinion. While T<sub>B2</sub> agreed with T<sub>A</sub> that language was mainly the medium, he added that content and language were inextricably linked through this relationship:

8	English translation	Original quote
T <sub>B2</sub>	but language conveys contents and without language, I can't convey these contents [...] <b>consequently, I can never split content and language completely.</b>	<i>aber Sprache transportiert Inhalte und ohne Sprache trans-, kann ich diese Inhalte nicht transportieren [...] also kann ich Sprache und Inhalt nie gänzlich voneinander trennen.</i>

T<sub>B2</sub> further explained that a strict division of content and language was also unattainable in terms of assessment despite current policies that suggest ignoring language in assessment:

9	English translation	Original quote
T <sub>B2</sub>	the moment when learners can't read and comprehend the task [...] their grades will always be affected. Therefore, <b>the claim that language would not be considered in assessment is simply wrong</b> [...] Maybe one shouldn't assess the number of comma errors in a final test, okay.	<i>in dem Moment, wo ein Schüler keine An-, keine Anweisung lesen kann und verstehen kann [...] wird sich das immer im Endeffekt auf die Note auswirken. Da ist also die Argumentation, dass Sprache in keine Beurteilung reinkommt, ist schlicht und ergreifend falsch [...] Dass es jetzt nicht in die Beurteilung reinfließen sollte, wie viel Beistrichfehler er vielleicht beim abschließenden Test gemacht hat, okay.</i>

In other words, language, for him, always played a role in assessment even though school policies did not reflect this. What this teacher seemed to favour, though, was meaning-based assessment rather than focusing on form. Due to these reasons, this teacher considered language to be a legitimate part of his teaching, meaning that, at least in principle, he would like to prepare content from a language-didactic perspective too, unlike T<sub>A</sub>. In his interview, T<sub>B2</sub> singled out one thing he found extremely important for his subject, namely precision of expression in either language. He reported that he taught this aspect explicitly and provided feedback on the learners' use of language in this regard, including recasts, asking the learners to rephrase, forming sentences together, and, more generally, raising awareness concerning the importance of precise language. He commented that the students hated this type of feedback and these activities, but he would do them nonetheless. Furthermore, he also said that he considered this aspect in his assessment for the following reasons:

10	English translation	Original quote
T <sub>B2</sub>	Well, in my subject, there are a number of things for which you need to express yourself relatively clearly sometimes. For instance, when given statistics or a graph and <b>I don't answer in a precise way, then it's simply wrong</b> . And often, the student would argue 'well this is what I meant', but he has to be able to indeed express it like that. But this also means that there are no excuses since I did support them [in this regard].	<i>Naja, in meinem Fach gibts schon einige Dinge, wo man sich sprachlich relativ klar manchmal auch ausdrücken muss. Also, wenn man bei einer Statistik oder Grafik, äh, nicht präzise genug antwortet, dann ist es schlicht und ergreifend falsch. Und da kann der Schüler dann schon oft argumentieren, naja, aber das hab ich eh gemeint, aber dann muss er in der Sprache in der Lage sein, dass auch sich tatsächlich so auszudrücken. Aber das heißt, es gibt dann keine Ausreden, wenn's dann nicht funktioniert, weil dann hab ich ihm die Hilfestellungen gegeben.</i>

Apart from this, he also mentioned explicit teaching of both English and German vocabulary, which sometimes also included the difference between German and English terms, but he added that this was not a focus in his teaching. Concerning considerations in terms of implicit language learning, he argued that deliberate and targeted overload worked best:

11	English translation	Original quote
T <sub>B2</sub>	Language input always works best if I <b>overwhelm the students purposefully</b> or only in one very little aspect.	<i>Sprachliche Inputs funktionieren ja dann am besten, wenn ich, wenn ich Schüler gezielt überfordere, oder überfordern nur in einem ganz, ganz kleinen Bereich.</i>

This extract, which seems to be inspired by Krashen's (1987) 'i+1' input hypothesis, shows that this content teacher keeps theoretical considerations concerning language in mind. Yet, he added that this premise was difficult to put into practice, as he lacked both materials and time to prepare

accordingly. Furthermore, he maintained that learner levels were difficult to determine and also varied quite extensively in most classes.

Finally, he said that he would like to include more writing in his subjects; a point he also addressed in the in-service teacher seminars he taught, but, so far, he had hardly managed to include writing longer texts in his own classes. He reported, however, that there was some shorter writing in his history classes also because the setting of laptop classes<sup>35</sup> facilitated this. Frequently, the students would get their tasks digitally and then complete them in written form, often having to hand them in online.

#### 7.1.1.2 Materials: selection, adaptation, and creation

Both teachers found the possibility of using historical sources, input materials, as well as didactic materials<sup>36</sup> in English extremely beneficial. They felt that this enriched their teaching because sources in their original language would also convey culture much more clearly, allowing for a more authentic approach. That being said, T<sub>B2</sub> argued that the price of using (historical) sources and materials in English was the extra amount of time needed for teaching CLIL units. As a consequence, other topics would have to be left out or could only be done rather briefly. Yet, he subsequently put this disadvantage into perspective by explaining that in the history curriculum, it was easier than in other subjects to set thematic priorities flexibly, as there was no standardized final exam for this subject. Furthermore, he argued that by taking more time when doing CLIL, the students experienced these topics more intensively while also improving their language skills, as emphasized in this extract:

12	English translation	Original quote
T <sub>B2</sub>	From the students' perspective, I find it very, very positive. [...] If I have to shorten one or the other topic, the students won't mind if, in return, <b>they can improve their English and experience the content in two languages more intensively, of course.</b>	<i>Aus Schülerinnenperspektive find ich's sehr, sehr positiv. [...] wenn ich den einen oder anderen Punkt etwas kürze, ähm, wird das wahrscheinlich auch für den Schüler und Schülerin nicht so schlimm sein, wenn sie dadurch [...] das Englische verbessern und vielleicht das Durchgemachte dann oft auch in zwei Sprachen natürlich auch intensiver erleben.</i>

From the teachers' perspective, this, however, means increased time for planning and preparing lessons since there are few appropriate CLIL materials available:

13	English translation	Original quote
T <sub>B2</sub>	<b>Those [= appropriate CLIL materials] are, I find, very, very rare.</b>	<i>Die [= angemessene CLIL Materialien] finden sich, find ich, sehr, sehr wenige.</i>

Consequently, the teachers said that they created a great amount of materials themselves. Especially, T<sub>A</sub>, who had been teaching bilingually for almost two decades at the time, stressed this:

<sup>35</sup> In school B, in some classes, all students are equipped with laptops, which they bring to school every day.

<sup>36</sup> The term *input materials* usually refers to informative texts or videos used to convey declarative knowledge. The label *didactic materials*, on the other hand, points to tools and tasks or input that has been didactically prepared.



14	English translation	Original quote
T <sub>A</sub>	<b>A lot!</b> [...] In the beginning, there was nothing available. We didn't even have an English textbook. <b>I basically made everything myself.</b>	<i>Viel [...] Vor allem am Anfang war ja nichts da. Da haben wir nicht einmal ein englischsprachiges Schulbuch gehabt. Da habe ich mir quasi <u>alles</u> selbst gemacht</i>

Especially when she started teaching, there was no material to draw from, meaning she created her own repository, which she continuously adapted and expanded. Yet, she pointed out that a bilingual setting opened up many possibilities in terms of source and material selection, which made this process, albeit necessary, easier and, in the end, more enjoyable. Here, T<sub>B2</sub> disagreed. For him, using materials in an additional language rather complicates lesson preparation because it requires more effort and time to select and prepare input materials from a language-didactic perspective; time that he sometimes simply does not have:

15	English translation	Original quote
T <sub>B2</sub>	I think I use a lot of English input materials and sources <b>but to prepare it also from a language-didactic perspective, I simply lack the resources and the time sometimes.</b>	<i>Ich setze glaube viele englische Materialien und Quellen ein, aber das sprachlich gut aufarbeiten, da fehlt mir manchmal die Unterlagen beziehungsweise die Zeit.</i>

Another factor could be that T<sub>B2</sub>, unlike T<sub>A</sub>, is not a language teacher and thus might be less familiar with the process of selecting, preparing, and working with English materials. Another reason could be that when it comes to language-didactic preparation, T<sub>A</sub> did not seem to have thought about it as much and therefore required less effort or time. This would also tie in with her statement that, for her, language was simply the medium of instruction, and nothing more, as mentioned in subsection 7.1.1.1.

T<sub>B2</sub>, in contrast, explicitly lamented the lack of CLIL materials that considered both content and language. So, for practical reasons, he often made use of EMI materials rather than CLIL materials despite his awareness of the importance of language-didactic considerations, concomitant with his background in language-sensitive teaching:

16	English translation	Original quote
T <sub>B2</sub>	we have quite a bit of bilingual materials, which can easily be just copied. <b>But I don't think that they are planned perfectly in terms of language acquisition</b> , but they offer simplified language and vocabulary definitions [...] it's this typical [...] <b>history in English</b> , meaning no typical CLIL materials [...] <b>One can use these [= EMI/ bilingual materials] easily and readily.</b> But they don't really promote language acquisition. There are <b>hardly any materials with a proper CLIL approach.</b>	<i>wir haben einiges an bilingual Materialien, wo man sehr leicht Kopien herausnehmen kann, die sind finde ich jetzt nicht perfekt gestaltet was den Spracherwerb anbelangt, aber sie bieten meistens eine vereinfachte Sprache, plus Vokabelerklärung [...] so dieses klassische [...] Geschichte auf Englisch, das heißt also keine, keine klassischen CLIL Materialien [...] die setzt man sehr einfach und sehr gerne ein. Ähm, so wirklich fördern für den Spracherwerb, dass jetzt diese Übungen in, in eine wirkliche CLIL Ausrichtung gehen, finde ich, gibts recht wenig Materialien.</i>

In short, the teachers stressed the need for new and appropriate materials throughout the interviews. For this reason, amongst others, the teachers appeared keen to participate in this study.

In terms of the procedure of material design, the two teachers seemed to follow different approaches. T<sub>A</sub> usually thought first about the didactic method she would like to employ, then she looked for sources and input materials as well as ready-made didactic materials, which she normally adapted, or she developed her own tasks accompanying historical sources or input resources. T<sub>B2</sub> tended to do it the other way around, taking historical sources and input materials as starting points. However, he felt that selecting appropriate input both in terms of content and language level was very difficult and that the resources found almost always needed adaption. Once he had appropriate input and sources, coming up with tasks was easy for him, as he said he often used default tasks to save time. Nonetheless, he reported that he employed creative approaches on a regular basis to cater for different learner styles and to keep up motivation. As for principles the teachers both intend to consider, they mentioned student-centredness and having a variety of engaging approaches, tasks, and materials.

### 7.1.1.3 Learner needs: subject and language learning

In the initial interviews, learner needs were discussed, and the points mentioned were often reflected in the teacher's reported classroom practices and materials outlined above.

In terms of content-related learner needs, T<sub>A</sub> only stressed lack of interest and motivation concerning history and politics, which, from her point of view, was not connected to language. Despite this view, she still hoped that participating in this research project with its intention to work more on subject-specific language could also contribute to higher levels of student motivation. This is something that T<sub>B2</sub> seemed to agree with, as for him, general affective issues were closely connected to language. According to T<sub>B2</sub>, the most pressing issue related to content learning was low frustration tolerance and lack of focus, meaning that learners gave up easily and were not willing or able to deal with something for a longer time or more than once:

17	English translation	Original quote
T <sub>B2</sub>	<b>Many have relatively low frustration levels,</b> I find. This means that they give up easily when doing tasks involving language [...] One rather needs to train them to deal with a source more deeply, to be willing to look more closely. [...] <b>This is not only related to language, but it concerns general willingness</b> to do something more often.	<i>Sehr viele haben, haben ein, ein recht niedriges Frustrationslevel find ich, das heißt sie haben, sie geben bei sprachlichen Aufgaben relativ schnell auf. [...] dass man sie eher dafür hintrainieren muss, dass sie sich länger mit einer Quelle auseinandersetzen, dass sie bereit sind, etwas genauer sich anzuschauen. [...] das ist jetzt nicht nur sprachlich, sondern da gehts allgemein um, äh, um eine Bereitschaft etwas öfters zu machen.</i>

He further explained that this could be connected to language since language barriers could be the reason for their struggles:

18	English translation	Original quote
T <sub>B2</sub>	I think, very often, <b>it goes hand in hand that one has language issues and little willingness to deal with difficult topics at the same time.</b> Very often, one leads to the other, so after a few sentences, I can't keep up and my brain doesn't even try to make an effort.	<i>Ich find halt oft gehts Hand in Hand, dass man sagt, ähm sprachliche Probleme und gleichzeitig wenig Bereitschaft sich mit einem schwierigen Thema auseinanderzusetzen. Oft führt das eine zum anderen, ich steig nach ein paar Sätzen aus und dann versuch ich das Hirn erst gar nicht anzustrengen.</i>

To address these issues, he suggested continuously practising endurance and patience in class with the help of reading strategies, such as pushing them to repeat the same procedure again and again, each time looking at different aspects from different perspectives to make sure they process the content deeply. Apart from that, T<sub>B2</sub> could not think of any other subject-specific learner needs, as he argued that his subjects were not difficult in the narrower sense. To put it differently, he thought that the level of difficulty of the content was usually not the reason why someone struggled. Instead, he listed some more language-related problematic areas, such as the use of imprecise and vague language or the lack of general academic and subject-specific vocabulary. T<sub>A</sub> also highlighted gaps in vocabulary as the main language-related learner need:

19	English translation	Original quote
T <sub>A</sub>	In the beginning, it's difficult for them to get into the use of English <b>because they have problems with vocabulary</b> [...] It takes a while, but as soon as <b>they get used to it</b> , I don't see it as a disadvantage anymore.	<i>Es ist am Anfang schwierig für die Schüler reinzukommen ins Englische, weil es Probleme mit dem Vokabular gibt [...] Es dauert ein bisschen und sobald sie aber drinnen sind, ähm sehe ich das nicht als Nachteil.</i>

While acknowledging the students' struggle with vocabulary, she assumed that issues with vocabulary tended to become less pronounced throughout their school career all by itself. She also reported that some of her colleagues worked with vocabulary lists, but to her, these lists were not a viable solution, as she preferred a more implicit approach to vocabulary acquisition. In general, she believed that BE was very beneficial for the language development of their learners:

20	English translation	Original quote
T <sub>A</sub>	<b>The students' language skills benefit immensely.</b> The bilingual ones <u>have no</u> problems when doing their standardized final [English] exams, written or oral. That's no problem for them at all.	<i>[Da] profitieren die Schüler was den Spracherwerb betrifft <u>total</u> davon. Das merkt man auch bei der Matura. Die bilingualen haben <u>überhaupt kein</u> Problem mit der Zentralmatura ob schriftlich oder mündlich. Das ist kein Problem für die.</i>

Turning to CDF-related learner needs, both teachers agreed that students often struggled with responding to performative verbs accurately. In Austria, the introduction of partly-standardized final exams has led to a widespread use of performative verbs for task and test design ("Operatoren"). Despite their application in all subjects, learners very often do not know how to deal with these tasks appropriately, according to these two teachers. They reported that learners often seemed to misunderstand, confuse, or misapply some of the performative verbs:

21	English translation	Original quote
T <sub>A</sub>	I think that <b>they simply don't know what to associate with performative verbs</b> [...] If it says 'explain, describe', they don't know what to do.	<i>ich glaube, dass sie nicht wissen, weil sie können einfach nix anmachen äh nix äh verbinden mit den Operatoren [...] Wenn da steht erkläre, beschreibe, dann wissen sie nicht was sie tun sollen.</i>

For T<sub>A</sub>, performative verbs are at the interface of subject and language learner needs. When asked whether subject and content learning needs could be connected, she responded:

22	English translation	Original quote
T <sub>A</sub>	<p>TA: Yes, <b>due to standardized testing and performative verbs, respectively, for sure, by now.</b></p> <p>R: So something external?</p> <p>TA: Yes, rather. Well, earlier [...] I would have said no, would have seen this entirely separate, but by now, no, <b>because they need to know what to do and have the linguistic tools to do so, respectively.</b></p>	<p>TA: Ja, aufgrund der Zentralmatura beziehungsweise der Operatoren auf alle Fälle mittlerweile.</p> <p>R: Also eher was von außen?</p> <p>TA: Ja, schon. Also früher [...] hätte ich gesagt nein, hätte das total separat gesehen, aber mittlerweile, nein, weil sie müssen ja wissen, was sie zu tun haben beziehungsweise müssen sie auch das sprachliche Werkzeug haben, um das auszudrücken.</p>

Earlier in this interview, she seemed to avoid questions concerning the connection or separation of content and language by reporting policies and previous practices. When it came to performative verbs, however, she clearly seemed to think that these linked content and language. Since this topic was addressed rather at the end of the interview, one interpretation could be that after talking about the different aspects of CLIL for quite some time, she might have reconsidered her position, deciding to answer this question directly now. Another interpretation could be that earlier she understood these questions differently, and now she indeed voiced her reflections rather than her practices.

Turning to potential reasons for the students' issues with performative verbs, both teachers agreed that one reason for the students' confusion was the way other teachers dealt with performative verbs:

23	English translation	Original quote
T <sub>B2</sub>	<p><b>the different subjects don't use the same performative verbs.</b> [...] On the one hand, we want these performative verbs, and on the other hand, we can't even agree on a common list. [...] <b>And I'm not even sure that all colleagues fully understand all performative verbs themselves.</b></p>	<p><i>die einzelnen Fächer haben keine gemeinsamen Operatoren [...] einerseits wollen wir von ihnen diese Operatoren, andererseits sind wir nicht einmal in der Lage, dass wir gemeinsam Operatoren überhaupt festlegen. [...] Und ich bin auch nicht so ganz sicher, dass alle Kollegen und Kolleginnen alle Operatoren gänzlich, gänzlich verstehen.</i></p>
T <sub>A</sub>	<p>Everybody needs these performative verbs. <b>Yet, nobody teaches or shows them how to apply the performative verbs.</b></p>	<p><i>Alle brauchen ja die Operatoren. Aber offensichtlich unterrichtet niemand, oder zeigt ihnen niemand, wie die Operatoren anzuwenden sind.</i></p>

For these reasons, the two teachers would not blame the students if they performed the wrong language function. T<sub>B2</sub> added that these issues seemed more serious in English than in German since, from his point of view, the German performative verbs were much more established than the English ones and thus more transparent and doable for the students. To address these issues, the two teachers called for more explicit and uniform teaching of these language functions. T<sub>B2</sub> suggested that one way to react to this problem would be to focus on a short list of central discourse functions, which should be explicitly discussed and practised in class. Apart from these practical and rather general suggestions, the teachers did not provide or evaluate concrete didactic solutions. This indicates that, even though they both seemed to be very aware of the issue, they had not developed anything tangible to approach this problem more systematically.

#### 7.1.1.4 Summary and implications for the design

To summarize, the teachers' initial position towards BE seemed very positive, talking lengthily about beneficial aspects and often qualifying negative experiences. Of course, this is not very surprising considering that they volunteered to participate in this research project. Nevertheless, positive attitudes towards CLIL and this study might be an advantageous foundation for the development and faithful implementation of new didactic material (see Dijkstra et al., 2017).

Furthermore, the above subsections have shown that the two participating teachers viewed the role of language somewhat differently and, considering their additional subjects, contrary to that what one would expect. While the language teacher ( $T_A$ ) tended to exclude language-didactic considerations, which corresponds to other studies in the Austrian context (e.g., Dalton-Puffer, 2007; Gierlinger, 2021; Hüttner et al., 2013), the content teacher ( $T_{B2}$ ) seemed ready to provide more room for that. However, it appeared that  $T_A$  started to reconsider her position at the end of the interview. Furthermore,  $T_{B2}$ 's answers indicate that the actual realization of language-sensitive content teaching can be rather challenging. A discrepancy between beliefs about successful CLIL and actual teacher practices was also reported by van Kampen et al. (2018) or Milla and García Mayo (2021), suggesting that teachers do not just need awareness for certain issues but also better support in implementing strategies for effective CLIL instruction.

Difficulty in implementation mainly seems to stem from the lack of appropriate materials and dissatisfaction with the current supply, according to the two interviewees. As a consequence, they often created their own materials or, due to lack of time, used material that they knew was not ideal from a language-didactic point of view. Since issues related to materials were repeatedly brought up, it can be assumed that this was the teachers' biggest concern in terms of BE, which might be a reason why the teachers seemed to welcome this research project. This result ties in with a number of studies, such as Ball (2018), Hahn (2019), Meyer et al. (2015) or Morton (2013), who all reported a pressing need for appropriate CLIL materials.

Turning to learner needs, both interviewees mentioned general issues, namely low frustration tolerance and lack of motivation, interest, or focus. These issues are relevant for history learning but, of course, also for many other subjects (see Otwinowska & Foryś, 2017, who reported similar findings for CLIL mathematics and science). These are aspects that should be kept in mind when designing the interventions. In line with Somers and Llinares (2018), who found that low language proficiency impedes motivation to learn academic content in a CLIL setting, it is hoped that a scaffolded approach that supports linguistically weak learners could promote student motivation towards content learning too. By creating small doable steps and providing linguistic support measures, materials could better cater to learners with short attention spans. Additionally, such a scaffolded and language-based approach might encourage and enable these learners to process content more deeply. This seems supported by the interview data, as these general affective issues were indeed said to be connected to language.

As for language-specific learner needs, both teachers mentioned lack of vocabulary and one teacher stressed the need to practise precise language. Again, these are points that should be considered in the intervention. Finally, both practitioners reported that students struggled with performative verbs, which was also observed in a study by Dalton-Puffer and Bauer-Marschallinger (2019). In this view, the interviewees agreed that working on CDFs, which are encapsulated by these verbs, could address a very pressing need of their students. Consequently, a CDF-based approach appeared to be a viable option for designing an intervention for these groups.

### 7.1.2 Focus group interviews with the students

The focus group interviews with students intended to shed light onto the students' perspectives on and their previous experiences with BE, as well as their wishes for the future in this regard, ultimately aiming to get to know the context thoroughly but also to tailor the interventions to the learners' needs. In other words, the students' perspectives and experiences were taken seriously to facilitate the viability and success of the didactic design, responding to calls for more student involvement in educational research, e.g., by Coyle (2013) or Groundwater-Smith and Mockler, (2016). The interview with five students (two male, three female) of group A took place in January 2019 prior to the start of the first intervention. Group B also consisted of five students (two male, three female) who were interviewed after the completion of cycle 1 and before the start of their own intervention (cycle 2) in March 2019. The results of the QCA of these two interviews are structured in five subsections, namely *views on BE* (7.1.2.1), *pedagogical practices related to language learning* (7.1.2.2), *bilingual materials* (7.1.2.3), and *learner needs* (7.1.2.4), concluding with a *summary and implications for the design* (7.1.2.5).

#### 7.1.2.1 Views on BE and previous experiences: benefits and problems

To begin with, the prevailing mood of the conversation was profoundly different in the two initial interviews with the students. While group A seemed rather positive and optimistic about their programme, mentioning a great number of benefits, group B appeared dissatisfied and pessimistic, listing mostly problems and drawbacks. The only benefit mentioned by both groups was the programme's positive impact on language learning, especially in terms of vocabulary:

24	English translation	Original quote
ARJ01 (A)	when you're talking to people the same age who are only taught in one language, <b>you realize that you have an advantage in English, well in terms of vocabulary.</b>	<i>wenn man jetzt mit ähm gleichaltrigen Menschen spricht, die einfach nur jetzt auf, in einer Sprache den Unterricht haben, dass wir dann schon ein Vorteil haben im Englischen, vokabeltechnisch halt.</i>
ICK01 (B)	the advantages <b>one takes away more</b> , I think, concern English. <b>It often happens to me that I can only think of English terms</b> and I can't even think of the German ones anymore.	<i>die Vorteile, man nimmt halt schon glaub ich mehr mit, dann auf Englisch. Also mir fallen auch meistens, also fast nur englische Begriffe dann ein und die deutschen Begriffe dann überhaupt nicht mehr.</i>

Interestingly, this is the only advantage mentioned in the interview with group B and even this one has a slightly negative undertone, implying that their German might suffer. Additionally,

another student of group B disagreed with his colleague and maintained that their language skills did not benefit from BE:

25	English translation	Original quote
IMJ07 (B)	Well, <b>I don't think that [being in] a CLIL class benefits your English skills whatsoever.</b>	<i>Also, ich glaub nicht, dass diese CLIL Klasse irgendwas bei deinen Englisch skills weiterbringt.</i>

He explained that, ultimately, their language skills could not improve since they were not doing 'real' CLIL in the first place:

26	English translation	Original quote
IMJ07 (B)	<b>I don't even notice that I'm in a CLIL group</b> because the only thing that could happen would be that the maths teacher maybe entered the room saying 'good morning' in English.	<i>ich merk nicht mal, dass ich in einer CLIL Klasse bin, weil das einzige was höchstens passiert ist, dass die Mathelehrerin vielleicht mal reinkommt und auf Englisch 'good morning' sagt.</i>

From this extract, it seems that group B had not really received much purposeful instruction in English and thus could not think of any benefits of the approach. Group A, on the other hand, had experienced more contact time with English and, consequently, listed a variety of perceived advantages for their language skills but also for their future plans and general quality of their education. To be more precise, apart from better vocabulary, they felt that their programme was beneficial for their fluency, their accent, subject literacies, as well as their bilingual proficiency:

27	English translation	Original quote
ARJ01 (A)	It's not a disadvantage at all because <b>we are constantly animated to think in both languages</b> because we are supposed to answer in the language the teacher [...] addresses us with.	<i>Es is gar kein Nachteil für uns, weil wir einfach durchgehend animiert sind nachzudenken auf beiden Sprachen, weil wir müssen antworten auf die Sprache, die uns der Lehrer [...] gerade fragt.</i>
OPB04 (A)	one advantage is that you can speak extr-, well, <b>very fluently</b> [...] and you also speak it more naturally.	<i>ein Vorteil ist es, dass man extr- also sehr flüssig dann Englisch sprechen kann [...] Und man spricht's natürlicher.</i>
ARJ01 (A)	often, in normal English [class] [...] there is only small talk, it doesn't really dive into subject matter. And now, for instance, we know, <b>I know how to explain historic events in English, using the right terminology</b> [...] and one can imitate the <b>accent</b> a bit.	<i>oftmals das normale Englisch [...] ist meistens ja small talk, also es ist nicht so in Thematik und jetzt haben wir halt zum Beispiel (.) wissen wir, ich weiß nicht also geschichtliche Ereignisse können wir auch auf Englisch mit den Fachvokabeln einfach erklären [...] und man schaut so ein bisschen sich den Akzent ab</i>

The first quote also points towards the students' impression of a good quality of education since it seems more cognitively stimulating. In this respect, they also explained that the bilingual programme could offer better projects as a result of their high English proficiency:

28	English translation	Original quote
IKS12 (A)	We have the <b>better projects</b> because we also speak English and (.) well more English [than the mainstream programme].	<i>wir haben viel bessere Projekte, dadurch dass wir auch Englisch sprechen und (.) also viel mehr Englisch sprechen.</i>

Furthermore, the respondents described their lessons as more engaging due to several reasons:

29	English translation	Original quote
OPB04 (A)	There's <b>variation</b> when you are not only [using] the same language.	<i>Es ist abwechselnd, dass man nicht nur die ganze Zeit dieselbe Sprache.</i>
SAA03 (A)	It's also <b>exciting</b> because especially in history, we've covered most of it [= the topics] already in lower secondary and now, if you hear it in another language, then it's more <b>interesting</b> .	<i>Es ist auch spannend, weil vor allem von Geschichte haben wir den Großteil in der Unterstufe ja schon jeder mal durchgemacht und jetzt hört man's halt auf einer anderen Sprache, das ist dann interessanter.</i>
IKS12 (A)	I'd say it's <b>curiosity</b> because there are many (.) English words I don't know [...] Then, when I hear a new word and the teacher tries to explain it in English, I'll remember it better than if I only heard it in German.	<i>Ich würd sagen, es ist sehr Neugier, weil ich einfach viele (.) Wörter im Englischen nicht kenn [...] Dann wenn ein neues Wort auftaucht und die Professorin versucht uns gerade aufzuklären in Englisch, dann bleibt das mehr hängen als wenn ich das deutsche Wort einfach nur höre.</i>

As a result of the engaging nature of BE, the students of group A felt that the additional language was an enrichment rather than a burden, helping them to stay focused and motivated, which seems especially important since the Austrian history curriculum tends to be repetitive in terms of historical eras and topics covered in lower and upper secondary.

Finally, the students also mentioned a number of advantages for their future plans, which for some were the decisive factors for joining the bilingual programme. In other words, they seemed to be aware of the current status of English as lingua franca and the possibilities connected to high-level English and multilingual proficiency, which could allow them to go abroad after school, including their countries of origin:

30	English translation	Original quote
OPB04 (A)	I'd like to <b>go abroad</b> after school, and I thought <b>English would be very useful then</b> .	<i>Ich würd nach der Schule dann gern in Ausland gehen und dachte da wär's auch ganz praktisch Englisch.</i>
ARJ01 (A)	because English is a world language, well, <b>the world language</b> , it is very advantageous.	<i>weil Englisch halt jetzt eine Weltsprache, also <u>die</u> Weltsprache ist, ist es einfach vorteilhaft.</i>
HIP11 (A)	after school, one can just <b>go and work anywhere</b> , which is a huge advantage, especially if you are not from Austria and you would like to go back to your <b>home country</b> , so to speak.	<i>man kann nach der Schule eigentlich überall hinfahren und arbeiten und es ist eigentlich ein großer Vorteil, ganz besonders, wenn man eben nicht aus Österreich kommt sondern sozusagen wieder in sein Herkunftsland zurückkehren möchte.</i>

Group A also talked briefly about some drawbacks of BE. First of all, they reported that the workload was very high, resulting in little free time:

31	English translation	Original quote
IKS12 (A)	this means <b>more work</b> , in fact, and <b>less free time</b> and <b>more stress</b> and that is a huge disadvantage.	<i>das heißt mehr Arbeit eigentlich und weniger Freizeit und mehr Stress und es hat schon diesen großen Nachteil.</i>

Secondly, they said that they sometimes struggled with expressing themselves in German:

32	English translation	Original quote
OPB04 (A)	one disadvantage would be for me that, sometimes, I just <b>forget the German words</b> and I only remember the English ones.	<i>ein Nachteil für mich wäre, ist, dass ich einfach die deutschen Wörter manchmal vergesse und nur noch das Englische weiß.</i>

This negative aspect was echoed in group B, linking it to concerns about assessment:



33	English translation	Original quote
ICK01 (B)	It is always hard <b>when you are doing content [...] in English but then you are tested in German.</b> This makes studying very hard because I have to sit down at home and translate.	<i>Ist halt immer schwer, wenn man den Stoff [...] auf Englisch macht, und der Test ist dann aber auf Deutsch; ist es halt immer schwer mit dem Lernen, weil dann muss man sich auch wirklich dann zuhause hinsetzen und das übersetzen.</i>
BF05 (B)	in the beginning, I wanted to graduate in science but now that we are doing so much in English, <b>I don't really know how to do it in German.</b>	<i>weil ich wollte am Anfang in Naturwissenschaften maturieren, nur dadurch, dass wir sehr viel mit Englisch machen, weiß ich jetzt nicht wie das, wie geht das jetzt dann auf Deutsch.</i>

It appears that these students are not well prepared to really work bilingually, resulting in feelings of overload. This might be connected to the school's policy that teachers are supposed to teach a minimum of 72 CLIL lessons a year, distributed over all subjects, but there is no plan in place how the teachers are supposed to fulfil their goal systematically as a team. Based on this interview, it seems that these learners as well as their teachers felt pressured by top-down policies, fearing that CLIL would overcomplicate instruction and slow them down, keeping them from reaching their curricular goals:

34	English translation	Original quote
EOS12 (B)	our maths teacher goes through the content in a very incomprehensible way, and on top, she <b>tries</b> doing it in English <b>so that even fewer people get it.</b>	<i>unsere Mathelehrerin machts so, dass sie einen Stoff durchgeht, ihn keiner versteht und sie's dann auf Englisch probiert und das dann noch weniger Leute verstehen.</i>
IMJ07 (B)	in most subjects where teachers <b>try</b> to implement it, <b>the students wouldn't even get it in German.</b>	<i>in den meisten Fächern, wo's die Lehrer versuchen umzusetzen, verstehen's die Schüler nicht mal auf Deutsch.</i>
EBF05 (B)	we all have <b>so many topics to cover</b> so that the teachers really have to rush through the content. <b>Therefore, they can't really cope with CLIL.</b>	<i>wir haben einfach alle einen hohen Stoffdruck, dass die Lehrer einfach viel Stoff durchbringen müssen. Und dadurch schaffen sie das mit dem CLIL irgendwie nicht</i>

In this interview, the students also talked about the teachers' perspective in this regard, reporting that their teachers struggled with the implementation of CLIL and implying that, in the end, some of their teachers failed:

35	English translation	Original quote
ICK01 (B)	It [=CLIL] is <b>very difficult to implement</b> for the teachers. Well, we only have few teachers who really incorporate CLIL on a regular basis.	<i>Es ist schwer umzusetzen auf jeden Fall für die Lehrer. Also es, wir haben sehr wenig Lehrer, die das wirklich regelmäßig im Unterricht CLIL einfließen lassen.</i>
EBF05 (B)	well, the teachers usually <b>try to somehow pull through with CLIL.</b>	<i>also die Lehrer versuchen meistens, das irgendwie durchzubringen, das mit dem CLIL</i>

From the way the students formulate their responses, i.e., using phrases like “try”, “pull through”, or “can't really cope”, one could deduce that from the students' point of view, the teachers lacked the pedagogical competence to really implement CLIL. Another reason why teachers might struggle with the implementation of CLIL relates to the teachers' language competence. When asked whether CLIL had the potential to improve the quality of instruction provided that the teachers had the pedagogical know-how for CLIL, one interviewee stated the following:

36	English translation	Original quote
EOS12 (B)	I think it can improve it, but <b>some teachers are just really bad at English.</b>	<i>Ich find es kann besser werden, aber manche Lehrer können so schlecht Englisch.</i>

In general, the students of group B seemed very dissatisfied with their bilingual programme, pointing towards various aspects in need of improvement. From the students' point of view, the implementation of this programme seemed inadequate, resulting in negative opinions about CLIL, which appears to be in stark contrast to group A's situation.

#### 7.1.2.2 Pedagogical practices in the bilingual classroom related to language

Considering their completely different views on BE tied to their previous experiences at their respective schools, the practices relating to language differed significantly.

Group B reported that English was only included by watching English videos, and even then, teachers and students partly used German for the discussion of the video afterwards. According to these learners, they had never used English for any tasks, source analyses, worksheets, or group work projects, meaning that there was hardly any productive use of English, as can be deduced from these extracts:

37	English translation	Original quote
OVD11 Sfx (B)	R: What else do you use English for? <b>OVD11: Nothing, right?</b> Sfx: Only videos, actually.	<i>R: hm, für was verwendet ihr sonst noch die englische Sprache? [...] OVD11: Gar nicht, oder? Sfx: Nur Videos eigentlich.</i>
EBF05 (B)	<b>Well, we do watch it in English, but then it will be explained partly in German.</b>	<i>Also wir schauen es, es zwar es auf Englisch an, aber erklärt wird es dann eben teils auf Deutsch.</i>
OVD11 Sfxx (B)	R: So you have never done <b>group work</b> in English? Sfxx: <b>No.</b> R: So you never really <b>use English when you talk to each other?</b> OVD11: <b>No.</b>	<i>R: Ähm, also ihr habt auch noch nie eine Gruppenarbeit oder so auf Englisch gemacht? Sfxn: Nein. R: Und sprecht untereinander eigentlich nie Englisch? OVD11: Nein.</i>

Interestingly, this differs quite significantly from their teacher's account, who indeed gave the impression that he used English occasionally, including some limited focus on language learning too. Yet, he also stressed repeatedly how difficult it was to implement CLIL, potentially resulting in little actual use of the approach. This coincides with the students' view that they were not doing CLIL in the first place. When asked whether they wished for more English in history class, the learners were quite hesitant or even rejected the idea completely:

38	English translation	Original quote
IMJ07 Sfx (B)	R: <b>And would you like more English in history class?</b> (...) Shaking heads. IMJ07: <b>No</b> Sfx: <b>I don't know.</b>	<i>R: Und hättet ihr gern mehr Englisch in Geschichte? (...) Schüttelnde Köpfe. IMJ07: Nein. Sfx: Ich weiß nicht</i>

Their argument for their rejection of more English in history class was based on the fact that tests were in German, implying that more English would complicate studying, as already mentioned in

in the previous subsection. However, they also stated that if they were not tested in German, then they would not mind more English instruction and English materials in history. Yet, they also said that one needs to make sure that the level of English would not be exceedingly difficult to ensure that they could still follow. Considering that group B apparently did not experience a lot of English, the students of this group reported that they never received feedback in terms of language use. They told an anecdote of a native speaker teacher who corrected them and provided linguistic feedback and they agreed that this was something they would like.

In group A, the situation seemed entirely different, as these learners indeed reported that they used English for all types of learning phases and task types. The learners of group A said that T<sub>A</sub> promoted the use of English by insisting on the use of English in whole-class discussions, especially if learners were hesitant to use the target language or displayed linguistic deficits. For group work, they said that they often mixed languages, yet it would very much depend on their partners, as some tended to avoid English:

39	English translation	Original quote
ARJ01 (A)	she addresses those in English <b>that actually need it the most.</b>	<i>sie spricht die dann auf Englisch meistens an, dies eigentlich nötiger haben.</i>
SAA03 (A)	then we have to answer in English and she doesn't accept any German answers [...] <b>It's good that one cannot just take the German way out,</b> one has to think more and (.) practise.	<i>dann müssen wir auch in Englisch antworten, halt, sie, ähm, sie nimmt dann auch keine deutschen Antworten an [...] das ist gut, dass man also nicht den deutschen Ausweg nehmen kann, wenn man mehr nachdenkt und (.) übt.</i>
OPB04 Sfx (A)	R: If you are doing a topic in English, do you use English to talk to each other or do you rather use German? Or a mixture? Sfx: Mixture. OPB04: <b>=Mixture, it depends on the people too.</b>	<i>R: wenn das Thema jetzt auf Englisch ist, sprecht ihr dann auch auf Englisch miteinander oder ist es doch Deutsch? Oder eine Mischung? Sfx: Mischung. OPB04: =Mischung, kommt auf die Leute drauf an auch.</i>

In general, they said that they would appreciate it if accuracy was not central in content subjects, so that participating could feel easy and free:

40	English translation	Original quote
SAA03 (A)	You'll be <b>freer if accuracy doesn't count as much.</b>	<i>Man wird freier, wenn man, wenn die Sa-, ähm die Satzrichtigkeit nicht so gewertet wird.</i>
IKS12 (A)	in T <sub>A</sub> 's classes, I <b>tend to hold back</b> and think five times before I speak.	<i>bei der Frau Professor [Name von TA] da halte ich mich zurück und denke fünfmal nach bevor ich irgendwas sag.</i>

However, some students also said that although overlooking accuracy indeed promoted active participation, they would still like to have their language output corrected, which, in fact, happened occasionally. To be more precise, they reported that they received feedback on accuracy in terms of pronunciation, grammar, vocabulary, and sometimes also style:

41	English translation	Original quote
IKS12 (A)	But if she really realizes that there's a <b>better [=more formal] word,</b> then <b>she'll say so and that's okay.</b>	<i>Aber wenn sie jetzt wirklich bemerkt, dass da ein besseres [=formeller] Wort gibt [...] dann sagt sie das auch und dann passt das auch.</i>

SAA03 (A)	If a word is <b>mispronounced</b> or if you are not sure, then she'll <b>prompt it</b> and we should repeat. Or if a sentence is <b>grammatically incorrect</b> with the effect that it doesn't make sense or it somehow irritates, then she'll ask whether you're sure [...] and wants you to <b>rephrase</b> the sentence or correct it.	<i>Wenn ein Vokabel falsch ausgesprochen wird oder wenn wir uns nicht sicher sind, dann sagt sie es vor und man soll es nachsagen. Oder wenn der Satz grammatikalisch nicht passt, und zwar so, dass es eigentlich keinen Sinn ergibt oder störend ist, dann fragt sie auch nach, ob man sich sicher ist [...] und will, dass der Satz neu gebildet wird oder verbessert wird.</i>
--------------	--	--

Later, the students added that T<sub>A</sub> did not only provide linguistic feedback but also topicalized grammar, vocabulary, and pronunciation during history lessons, e.g., when dealing with a difficult English text. This seems contrary to what T<sub>A</sub> reported in her interview, as she said that she broke the habit of correcting students and teaching language in content subjects when she started teaching in the bilingual programme. In a later interview, T<sub>A</sub> realized that the students' perceptions were right, as she unconsciously had started to correct them. She argued that this might be the case because she was also their English teacher and thus more invested in their English skills or participating in this project could have affected her practices subconsciously too.

Interestingly, when asked whether the students would like to focus more on language relevant for history, both groups were slightly confused by that question, as they could not imagine what that could look like at first. Once that was clarified, both groups seemed to be in favour of explicit attention to linguistic features relevant for historical discourse, as can be seen in extract 42:

42	English translation	Original quote
ARJ01 (A)	Ehm, yes actually, because I think we are in the bilingual programme to <b>shine in English</b> and (.) considering this, <b>we only do very little</b> [...] but I guess then the <b>majority couldn't cope anymore</b> .	<i>Ähm eigentlich schon, weil ich denk wir sind in einem bilingualen Zweig, um eben in Englisch noch zu glänzen (.) und dafür machen wir eigentlich schon eher wenig, aber [...] ich glaub die meisten würden dann nicht mehr mitkommen</i>

While stating that he would welcome more focus on English in history, student ARJ01 feared that the majority of his peers would be overwhelmed by that. Group B would only like explicit attention to language provided the teacher was capable of doing so.

### 7.1.2.3 Bilingual materials: status quo and wishes for the future

Due to the different approaches to bilingual teaching, the use of bilingual materials differs quite extensively between groups A and B. While group A reported that they usually made use of a mix of English summaries and tasks, group B said that they were only presented with English input materials in the form of videos while the remaining didactic materials and input resources were in German only. Both groups seemed rather happy with the current situation, but, quite naturally, for different reasons. Group B said they liked the videos but would not want English worksheets since, as already mentioned above, English material would further complicate studying for German tests, from their perspective. Group A, on the other hand, appreciated a number of characteristics of the materials used in class:

43	English translation	Original quote
ARJ01 (A)	our teacher actually really does a good job because she's <b>encouraging interaction</b> .	<i>die Frau Professor machts eigentlich sehr gut, weil sie viel Interaktion sucht.</i>
SAA (A)	Her teaching methods are <b>creative</b> , actually.	<i>Ihre Lehrmethoden sind scho-, kreativ eigentlich.</i>
HIP11 (A)	they [=materials] are <b>designed by the teacher</b> , and therefore they <b>fit the lessons</b> .	<i>die sind, ähm, von der Frau Professor selbst angestellt und daher auch an den Unterricht angepasst.</i>

When asked what they would wish for in future materials, group A called for creative, interactive, engaging tailor-made materials and preferred worksheets with tasks over summaries, as one student explained:

44	English translation	Original quote
IKS12 (A)	<b>even the best handout [=summary] won't do any good if you only hand it out and don't do anything with it.</b>	<i>auch ein gutes Handout bringt nichts, wenn man's einfach verteilt und dann nichts draus macht.</i>

Furthermore, they said that they would like to work with material that helped them express themselves better and provided them with a glossary. Group B stressed that they would appreciate truly bilingual materials, i.e., materials using two languages, to avoid the necessity of translating for the test. Additionally, they wished for multimodal input and variation of didactic tools, as argued in this extract:

45	English translation	Original quote
ICK01 (B)	Maybe some <b>visual</b> -, well <b>illustrations</b> , not just texts or just crosswords [...] maybe also <b>translated</b> or so. So that you already have it in German and English.	<i>Vielleicht bildlicher an-, also Darstellungen, nicht nur Text oder nur Kreuzworträtsel [...] vielleicht auch übersetzt oder so. Dass man das gleich von Deutsch auf Englisch hat.</i>

#### 7.1.2.4 Learner needs: subject and language learning

Similar to the teachers' interviews, the students also repeatedly reported that their biggest issue in terms of content learning was general lack of interest, motivation, and purpose:

46	English translation	Original quote
ARJ01 (A)	<b>We don't need that [=history]</b> , nobody needed that ten years ago, why do we still learn that?	<i>Das brauchen wir ja nicht, das braucht keiner mehr vor zehn Jahren, wieso lernen wir das noch?</i>

A solution put forward by group A was to connect topics more to the present and highlight their long-term relevance. Group B did not offer any strategies, as they regard interest and motivation as inherently personal and thus something that could not really be changed.

As for concrete subject-specific challenges, group A mentioned that connecting different topics and comparing was very difficult on a conceptual level and that they struggled with timelines and ordering events. Related to this, the students also admitted that they had huge gaps in background knowledge. In other words, non-linear, a-chronological approaches were challenging for them because the students seemed to miss the bigger picture.

One point mentioned by group B was that textual source analysis, unlike visual source analysis, was difficult for them, as argued by this student:

47	English translation	Original quote
ICK01 (B)	well, for me, analysing pictures and [stating] what I'm seeing is easy, but <b>if there are texts, then not really.</b>	<i>also für mich Bilder irgendetwas analysieren und was ich da eben drinnen sehe, ist einfach, aber wenn es so Texte gibt, dann eher nicht wirklich.</i>

Regarding this topic, group A added that text analysis was more difficult due to archaic language, intersecting now with language-related learner needs. Here, the students explained that step-by-step explanation, including the discussion of linguistic features and lexical items, usually helped them process. When asked whether they had problems with detecting underlying intentions of the writer or taking over their perspective, the learners negated this question.

Moving on to linguistic learner needs in a narrower sense, both groups said that they lacked general and subject-specific terminology and that they struggled with translating accurately. To address this, the students usually resort to Google Translate or ask for the teacher's help. Apart from that, group B could not point to any specific areas, but they agreed that they generally struggled with language in the subject history. While discussing the issue of expressing historical content and concepts clearly and fluently, Group A pointed out that the language level varied quite extensively in this group. One student shared the following:

48	English translation	Original quote
IKS12 (A)	If it's about history [...] <b>I can talk fluently in German, but in English, I start stammering</b> , I realized. But yes, in that case I just try to <b>define something somehow.</b>	<i>Wenn's um Geschichte geht [...] da kann ich einfach freireden auf Deutsch, und auf Englisch tu ich dann schon öfter stottern. Das hab ich bemerkt, aber ja, sonst versuch ich einfach irgendwie irgendetwas zu definieren.</i>

Interestingly, her strategy for when she starts floundering would be to “define something somehow”. This already points towards the students' imprecision or misconceptions concerning language functions. Defining per se has little to do with fluency, so in this extract, “define” seems to be used as interchangeable substitution for any language function.

In both groups, the students said that they knew what different performative verbs would refer to, in theory. Yet, when asked whether they could work well with them, this student said he particularly struggled in test situations:

49	English translation	Original quote
HIP11 (A)	HIP11: I do, for tasks but <b>not in tests</b> [...] R: What's the difference? HIP11: [...] I don't know, I think it depends on the question or <b>what you have to describe</b> , respectively.	<i>HIP11: Bei Arbeitsaufträgen schon, bei Tests nicht. [...] R: Was ist der Unterschied? HIP11: [...] ich weiß nicht, ich glaube es kommt auf die Frage an, beziehungsweise was man genau beschreiben muss.</i>

HIP11's explanation for this difference seemed rather vague and again indicates misconceptions relating to the function of performative verbs. To him, everything would be “describing”, which would entail providing only surface information. The reason why this student said he could work well with performative verbs in school exercises but not in tests could be that during lessons, wrong interpretations of these tasks tend to be ignored, while in tests, these could lead to loss of points. Students of both groups also said that, sometimes, one would need to interpret performative verbs, such as “analyse”, since different teachers understand them differently:

50	English translation	Original quote
OPB04 (A)	<b>Every teacher wants something else</b> although they all write the same [labels]. <b>But one has to know the teachers a little to know what they want.</b>	<i>Ja also es will eigentlich jeder Lehrer was anders, obwohl sie dasselbe hinschreiben. Aber man muss doch immer so ein bisschen die Lehrer kennen, was sie da wollen.</i>

One student then outlined her approach to this issue, which again shows that there is great uncertainty concerning performative verbs:

51	English translation	Original quote
IKS12 (A)	I'd say one should [...] just write down everything you remember and <b>then let the teacher decide what is right and what is wrong</b> (laughs).	<i>Ich würd sagen so man sollte [...] einfach das ganze Wissen drunter schreiben und dann sollte der Lehrer entscheiden was richtig und was falsch ist (lacht).</i>

Later, IKS12 added that she was aware that this could be considered a misinterpretation of the task, but she would do it anyhow because often she simply did not know what she was expected to do. The solution proposed by the students was to work more explicitly with language functions, step-by-step, providing more guidance, also in terms of content, i.e., by adding keywords directly in the prompt to limit the topics.

Finally, when asked about the interconnection of content and language learner needs, both groups affirmed a close connection between content and language. The students all seem to think that content and language cannot be separated completely because language is needed to understand the task and express what one knows:

52	English translation	Original quote
EBF05 (B)	The best example is maths. If you <b>don't understand the task</b> , you won't be able to calculate.	<i>Das beste Beispiel ist Mathe. Wenn man die Angabe nicht versteht, kann man es nicht rechnen</i>
OPB05 (A)	You also have to know the words, otherwise you can't write it down. Or say it.	<i>Man muss auch die Vokabeln können, sonst kann man ja das nicht hinschreiben. Oder halt sagen.</i>

In general, it seems that both groups were aware that content and language learning are somehow linked which has an effect on instruction, learning, and assessment.

#### 7.1.2.5 Summary and implications for the design

In short, the initial interviews with these two groups of students indicate that the two contexts of this study present two different points of departure. Group A seemed accustomed to both instruction in English as well as explicit focus on language in content subjects, whereas group B appeared to lack experience in terms of well-implemented CLIL, suggesting that the bilingual programme at school B might be somewhat underdeveloped. In connection to these differing experiences, the students' opinions of BE are almost diametrically opposed. Group A listed a great number of advantages of BE. Group B, in contrast, almost exclusively talked about drawbacks, mostly stemming from unprepared teachers and ill-conceived implementation strategies; a finding also reported in other studies (e.g., Banegas, 2012; Cabezuelo Gutierrez & Fernández, 2014; Pérez Cañado, 2016a). Consequently, group B initially rejected the idea of more English.

Nevertheless, they said that they would be open to using English in history class on the condition that teachers would be well prepared and capable.

Based on the different conceptions of BE in these two schools, their in-class experiences also differed quite extensively. Again, group A reported that they not only received English input and completed tasks in English, but they also received feedback on their oral production. Moreover, their history teacher topicalized linguistic aspects from time to time, for instance when dealing with difficult texts. Group B, on the other hand, used English only for audio-visual input in their history lessons. Thus, neither group seemed to have experience with CLIL materials in the narrower sense, i.e., materials that consider content and language in integration. As a consequence, the students should not be expected to be familiar with the approach of the intervention, meaning that the materials to be implemented require clear instructions and continuous guidance. Furthermore, the intervention should consider well-received features of their current materials as well as wishes for the future to ensure acceptance by the learners despite introducing an approach they were not used to at this point. These features and principles entail that input and didactic materials should be interactive, cognitively engaging, multi-modal, and didactically prepared, which seems to align with general recommendations by the CLIL research community (e.g., Ball et al., 2015; Banegas, 2017; Mehisto et al., 2009; Meyer, 2013; Pérez Cañado, 2018). The features mentioned by the learners of this study seemed to orientate towards affect and cognition, and interestingly, similar points were also put forward in a study by Coyle (2013), who asked British learners about strategies for successful CLIL.

Zooming in on history materials, the learners of this study asked for a chronological approach to history while still making connections to the present explicit. This wish for chronology is somewhat at odds with current theories of history education stressing the importance of working with *second order concepts* (see, e.g., Seixas, 2017, or subchapter 4.1). In this respect, one could help learners by adding timelines or graphic organizers. The learners' request to link history to the present tense, in contrast, very much reflects central notions of history didactics, most notably *historical consciousness* (e.g., Rüsen, 1983, 2004, see also chapter 4), and should thus be considered in didactic materials.

As for the learner needs to be addressed in the intervention, the interviews with the students revealed that both groups lacked vocabulary and especially struggled with archaic language often present in old texts. The learners of group B also feared that being taught in the L2 might negatively affect their subject-specific L1 literacy, which corresponds to the results reported by Gablasova (2014), who assessed Slovak CLIL learners both in their L1 and L2. For these reasons, adding glossaries could be beneficial. Furthermore, as the interviewees found processing historic texts challenging, scaffolding of input should prove advantageous. The students interviewed stated that expressing themselves clearly and fluently when dealing with historical content was difficult. Thus, learner output should be supported by providing linguistic scaffolding. Another problematic area identified in these interviews was the understanding of performative verbs and the performance of academic language functions. As assumed by their teachers, the students'



answers indicated that they tended to misunderstand and confuse academic language functions and thus struggled in exams. Furthermore, the students confirmed the teachers' observations that other teachers often did not apply performative verbs consistently and appropriately, reinforcing the learners' insecurities in this regard. Again, an explicit CDF-based approach might support the learners in this regard.

In summary, to ensure that both groups can profit from the planned intervention, it seems vital to keep in mind their different initial positions, both in terms of previous experiences and opinions on BE, entailing different expectations concerning the participation in this study. Although the contexts of the two groups were very dissimilar, they shared a lack of experience with language-based scaffolding and a truly content-and-language-integrative approach. Yet, it also appeared that these two groups could benefit from such an approach, as they seemed to struggle with processing linguistically complex input, expressing subject-specific content, and the concept of performative verbs encapsulating academic language functions. Finally, despite their many differences, both groups seemed amenable to such an approach, as they all agreed on the importance of language for content learning and the close relations between the two.

### 7.1.3 Initial competency-based written tasks

The initial competency-based written tasks were conducted to explore and identify learner needs while also setting a base line for future evaluations of the intervention. The students of both groups completed the tasks prior to the interventions on topics previously covered in history class, which were early high civilizations in the case of group A and exploration and colonialization of the Americas in group B. Details of these tasks and their administration can be found in section 5.4.2. In the following, results of both groups in terms of history-based and linguistic rating are presented and examined both qualitatively and quantitatively. Furthermore, potential connections between language- and content-related descriptors are discussed in an attempt to determine those areas that might play a central role for the success in the discipline and should consequently be given enough space in the interventions. In addition to the qualitative analyses, these connections are further investigated via correlation coefficients (Pearson's  $r$  for normally distributed data and Kendall's tau  $b$  ( $\tau_b$ ) for non-normal data).

Table 7 presents the results of all students completing the initial task (T1). Without going into great detail, it can be seen that on average, both groups achieved very similar results, with almost exactly the same average values for language and content (see section 5.5.3 for more details concerning the scales). Additionally, all average results can be assumed to be normally distributed, meaning that all distributions follow a similar curve with the majority of values clustered around the mean, not differing statistically significantly from normal distribution (see appendix [section II/B/](#) subfolders [1](#) & [2/](#) files [cycle1\(A\) NA statistics](#) and [cycle2\(B\) NA statistics](#) for normality tests, plots, and descriptive statistics). Consequently, it can be argued that both groups, despite their different experiences with BE, seem to be rather comparable in terms of initial levels of proficiency; an observation confirmed by  $t$ -tests that indicate no significant

differences (content:  $M_A = 1.83$ ,  $SD_A = 0.31$ ,  $M_B = 1.82$ ,  $SD_B = .47$ ,  $t(31.25) = 0.05$ ,  $p = .960$ ; language:  $M_A = 1.37$ ,  $SD_A = 0.43$ ,  $M_B = 1.35$ ,  $SD_B = 0.59$ ,  $t(32.41) = 0.135$ ,  $p = .894$ ) and very small effect sizes (content:  $d = 0.04$ ; language:  $d = 0.09$ ).

Table 7. Overview of results: T1

group A	content	language	ratio C:L	overall	group B	content	language	ratio C:L	overall
ARJ01*	1.73	1.33	1.30	1.53	ABS04	2.50	2.33	1.07	2.42
ATC04	2.20	1.50	1.47	1.85	AKM12	1.96	1.67	1.18	1.81
AVS07	1.73	1.17	1.49	1.45	APK08	2.20	1.33	1.65	1.77
ELF03	1.41	0.80	1.77	1.11	ARC11	1.07	1.00	1.07	1.03
ELH01	1.72	1.17	1.48	1.45	ARM03	2.27	2.17	1.05	2.22
ETS12	1.61	2.00	0.80	1.80	DRI04	1.19	0.60	1.98	0.89
EVA02	2.13	1.83	1.16	1.98	EBF05	1.93	1.20	1.60	1.56
EVS03	1.96	1.20	1.63	1.58	EOD03	2.33	2.17	1.08	2.25
HIP11	2.10	1.00	2.10	1.55	HRG10	1.86	1.00	1.86	1.43
ICM01	2.03	2.00	1.02	2.02	ICK01	1.24	0.67	1.86	0.95
IJT12	1.70	1.33	1.28	1.52	IMJ07	1.52	0.67	2.28	1.09
IKS12	1.96	1.50	1.31	1.73	LED08	1.61	1.67	0.96	1.64
LES02	2.00	1.67	1.20	1.83	NGS01	2.28	1.17	1.95	1.72
NNM05	1.12	0.67	1.68	0.89	OVD11	1.52	1.00	1.52	1.26
OPB04	1.48	0.80	1.85	1.14	UCQ07	2.60	2.50	1.04	2.55
ORH09	2.23	1.50	1.49	1.87	UKV05	1.69	1.17	1.45	1.43
SAA03	1.83	1.83	1.00	1.83	USN05	1.11	1.00	1.11	1.05
*ARJ01 did not complete T2. Thus, his scores are not considered in any statistical measures.					WAS01	2.00	1.50	1.33	1.75
					ZEA11	1.68	0.83	2.01	1.26
average	1.83	1.37	1.33	1.60	average	1.82	1.35	1.35	1.58
range	1.11	1.33	1.30	1.12	range	1.53	1.90	1.31	1.66

The main difference between the two groups appears to be the range of the results, manifesting in higher standard deviations in group B than in group A (see *t*-tests above). This means that group B is more heterogeneous, especially on the language scale. Comparing language and content results, all but three students of the combined data set ( $N = 35$ ) received higher scores on the content scale. Of course, these differences might stem from the design of the rating rubrics, which have not been calibrated to each other, but it nonetheless points towards a considerable need for improvement of their academic language skills. What is more, the data suggests that content and language results correlate (group A,  $n = 16$ ,  $r_p = .58$ ,  $p = .019$ ; group B,  $n = 19$ ,  $r_p = .81$ ,  $p < .001$ ; combined,  $N = 35$ ,  $r_p = .74$ ,  $p < .001$ ). In other words, by tendency, the better the language results, the better the content results and vice versa.

### 7.1.3.1 History-based rating results

Turning now to the content-related results in more detail, Table 8 on the next page shows that the two groups are very similar when it comes to individual descriptors<sup>37</sup>, with average values ranging

<sup>37</sup> See subsection 5.5.3.2 for more information on these descriptors and section 4.2.5 for their inherent concepts.

from 1.49 to 2.24. Overall, both groups seemed to do better on the level-related scales than in terms of historical competences.

Level-related rating: Especially the descriptor *target level* yielded rather high results. This means that most of the time, students tended to perform historical thinking skills on the intended level, e.g., reflecting when they were asked to reflect rather than re-organising or reproducing knowledge. Furthermore, the results suggest that, on average, the learners somewhat relied on accurate and relevant information and managed to present their declarative and procedural knowledge in a relatively systematic fashion. The example below exemplifies an answer mostly convincing in terms of level-related descriptors, with a level rating of 3-3-2/  $M = 2.7$ :

Table 8. Average history-based ratings

	average A (T1)	average B (T1)
<i>target level</i>	2.22	2.24
<i>accuracy/ relevance of content</i>	1.74	1.82
<i>systematicity</i>	1.74	1.79
<b>overall level stage</b>	1.91	1.95
<i>target competence</i>	2.25	1.99
<i>justification/ comprehensibility</i>	1.50	1.49
<i>scope of content</i>	1.56	1.58
<b>overall competence stage</b>	1.74	1.69
<b>overall</b>	1.83	1.82

53	EV: {Well it is a reminder of the oppression and the tragedy of the past, for us to learn. Since those were the stepping stones for racism and discrimination. That is why I think this painting is important for the generations living in the 21st century. It is definitely an important thing to consider when talking about oppression in a broad sense and History in a sense.} <sup>38</sup>
Item 5	
UCQ07 (B)	

When asked to judge the relevance of the source for the 21<sup>st</sup> century (item 5), the student clearly engaged in reflecting by stressing its significance for today's society. By pointing towards racism, he identified a central issue depicted in the source, which is still current, thus including relevant aspects that do not contradict established historical facts. As for *systematicity*, this answer is somewhat circular, with the more general and somewhat vague statement at the end. A different structure, for example from the general to the more specific, would make this answer more efficient. Relating to this, a more direct connection to the contents of the source would also improve the answer, e.g., by explaining how exactly this source depicts racism, oppression, and discrimination. This way, the justification for his claim would have been stronger.

Sample 54 provides an example for a low target-level answer, with a level rating of 0-2-x/  $M = 1$ :

54	RE: {In some countries slavery hasn't still been abolished.}
Item 5	
ICM01(A)	

Here, the student only reported a current issue, but she did not link it to the picture and its content, let alone reflect on the implications of this statement. In other words, she reproduced declarative

<sup>38</sup> CDFs are indicated with abbreviations (see List of abbreviations). Moreover, CDF-episodes are marked with { } while basic CDFs are identified with [ ]. Slashes between CDF abbreviations indicate that both codes apply. CDF codes in subscript, however, mark an alternative interpretation. Furthermore, learner samples were not edited for linguistic accuracy. Therefore, they contain the learners' original mistakes and errors.

knowledge instead of reflecting on the picture's relevance for the 21<sup>st</sup> century. The facts included are nonetheless accurate and also relevant for the task, yet she failed to explain how or why it might be relevant. *Systematicity* has not been rated for this example, as there is too little language to assess this aspect (which is marked with an x above).

Competence-related ratings: Samples 53 and 54 above share that they both lack a clear and explicit connection to the source, albeit to varying degrees. This is an issue that can be observed in the data to a great extent, manifesting in low average results for *justification* and *scope*. *Target-competence* results do not seem to be as affected since students, by and large, indeed managed to perform the intended competence, i.e., deconstructing the source (methodological competence) or establishing a connection to the present (orientation competence). Yet, very often, their observations, evaluations, explorations, etc., remained unjustified (*justification*), and the learners also seemed to struggle to cover all necessary parts and provide enough details for the individual tasks (*scope*). The following sample exemplifies a lack of *justification* and *scope* as defined in the rubrics used in this study:

55	RE: {Exploration was not always about exploring, but to find ways how you can export and import goods – The /Handelswege/ were also important to build new relations to an existing country – Many slaves were used for the colonialization – There were many inventions when the explorations began}
Item 4	
ZEA11(B)	

Item 4 required the students to assess the validity of the source. Yet, ZEA11 provided a number of facts she remembered from class without even mentioning the picture, which seems similar to IKS12's approach to let the teacher decide what is relevant and what is not, as mentioned in the interview with the students of group A (extract 51 on page 153). One explanation in ZEA11's case could be that she thought that by mentioning these bits of declarative knowledge somewhat related to the contents depicted, she would judge the source as valid. However, this would constitute a very indirect approach that leaves quite some room for interpretation and is not very tangible for readers. As such, this answer was rated as 1-1-1 on the competence rating scale, as (1) she did not deconstruct the source but included concepts related to the source, (2) the connection between source and answer is only comprehensible with great effort, and (3) while she did provide a number of correct facts, she still missed the main point.

In contrast to ZEA11, who did not give a verdict on the validity of the source, many students provided a definite decision but could not clearly justify their views. An example (rated as 1-1-1 on the competence scale) is provided below:

56	RE/EV: {Today we know it went different. [RE/EA: The main reason why the Europeans won were diseases and not their weapons or because they were mistaken for gods or something else. Furthermore, they tricked the aborigine to kill each other.]}
Item 4	
UKV05(B)	

In this example, the student negates the truthfulness of the source, but the reasons for his judgement are not fully comprehensible. On the one hand, UKV05 mainly reports long-term consequences while this source only illustrates an imagined first contact. Looking at the picture, it is not quite clear what the "winning" refers to and whether or how his report relates to the

validity of the source. Again, it could be understood as an implicit justification, as he outlined an alternative narrative related to the picture. Yet, another reason for the unsuccessfulness of this move is that his point about misplaced worship even somewhat contradicts his claim. In the source, the native inhabitants offer their treasures, which could be interpreted as showing their reverence. UKV05, however, used their supposed worship as an argument against the validity of the source; yet, following his logic, it would actually support the validity. As such, this sample illustrates two main points often observed in the data: First of all, learners struggled to make overt connections, e.g., by making explicit how their points contribute to their evaluations, thus clearly justifying their claims. Secondly, they often failed to comprehensibly base their answers on what was provided by the source. Yet, underpinning their assessment on the source alone does not suffice for a complete justification (i.e., achieving stage 3 on the *justification* scale) either, as sample 57 below demonstrates:

57	EV: {It shows a bit of the truth, because of [DS: the inhabitants running away from the boats in the background of the picture. It also shows, that most of the explorer were Catholics because of the man holding up the crus. But it doesn't show that, [RE: if the inhabitants didn't want to become a slave, the got killed by the explorer's men.]]}
Item 4	
NGS01(B)	

NGS01 sustained her assessment with what she saw or did not see in the picture. For a complete justification, she would have also needed to relate her observations to the historical context, i.e., assessing to what extent the depicted content corresponds to the historical context. Again, it seems that she implied such a connection. For these reasons, this sample was awarded stage 2 on the *justification* scale. Making these links more explicit would have supported the communicative purpose of this text, potentially yielding full points for *justification*.

In the example below, the student manages to include the picture and presumably the historical context too, but still could not use this information to clearly support her claim:

58	EV: {Yes, I think that it is a truthful description. [DS: You can exactly see that just one person is sitting and doing nothing and all others are hardworking][EA <sub>(RE)</sub> : because it was hierarchically structured. The leader was on top.]}
Item 4	
AVS07(A)	

Following her verdict, she first offers a description of the picture. Interestingly, she then continues with giving a reason for what she was seeing. It can be assumed that “because it was hierarchically structured. The leader was on top” refers to the society of Ancient Egypt, thus pointing to the historical context, but it could also just refer to what she saw in the picture. As the point of reference is not linguistically marked, the reader simply cannot know what she intended. Considering the prompt, one could of course presume that she included these parts to justify her claim. Still, the reader has to infer all these connections to accept this as a justified evaluation. Therefore, *justification* was rated as stage 1. Consequently, clearer reference to both picture and historical context as well as more precise linking would make her communicative intention clear, which can only be realized with linguistic skills not yet mastered by this student in this extract.

Limited *scope* seems to be connected to lacking or missing justification since crucial aspects were often missing, and concomitantly often only little detail was provided. Statistically, this is underlined by a strong, significant correlation between these two domains ( $\tau_b = .74, p < .001, N =$

35). In many cases, such answers appear to be superficial. Especially when asked to speculate about the artist's intention for producing the source, the students frequently provided surface answers, such as the one below, which was rated stage 1 in terms of *scope*:

59	EA: {To show other people who aren't there how the situation was.}
Item 2	
HRG10(B)	

This exploration of potential motives is very generic and not linked to the specifics of the source. Furthermore, it is not really an exploration as such, since HRG10 presented his answer in a very matter-of-fact style and did not provide alternative explanations as to why the artist produced exactly this image. There were also many cases of learners relating the intention of the artist to themselves or people nowadays, indicating that these learners often perceived historical sources from their contemporaneous context, as in the example below:

60	EA/EO: {I think he or she produced this illustration to show the old Egypt to the people nowadays, because Egypt played an important role for our history.}
Item 2	
IJT12(A)	

Here, IJT12 only considered the current generation as the target audience, completely ignoring the historical context of the source. As IJT12 missed the main point of a justified contextualization, this case was rated as 2-1-1 on the competence scale.

### 7.1.3.2 Linguistic rating results and their connections to content results

Moving on to the linguistic rating, the absence of average values above 2.0 is quite striking, meaning that there is definitely room for improvement on all scales (see Table 9). Overall, *appropriate use of hedging* ( $M_A = 1.00$ ,  $M_B = 1.11$ ) and *linking both in terms of form* ( $M_A = 1.25$ ,  $M_B = 1.05$ ) and *function* ( $M_A = 1.25$ ,  $M_B = 1.21$ ) proved to be especially problematic for these learners. This is reflected in a considerable number of

Table 9. Average linguistic ratings: T1

	average A (T1)	average B (T1)
<i>choice of CDF types</i>	1.75	1.95
<i>composition of CDF types</i>	1.46	1.44
<i>linking in terms of function</i>	1.25	1.21
<i>linking in terms of form</i>	1.25	1.05
<i>use of hedging</i>	1.00	1.11
<i>use of nominalisation</i>	1.63	1.42
<b>overall</b>	1.37	1.35

zero-level ratings, which were assigned ten times for *hedging* and seven times for *linking in terms of function* and *form* respectively, meaning the students made little to no (appropriate) use of these linguistic devices. Additionally, many learners struggled with logical *composition of CDF types* ( $M_A = 1.46$ ,  $M_B = 1.44$ ) and *appropriate use of nominalisation* ( $M_A = 1.63$ ,  $M_B = 1.42$ ). Concerning *composition of CDF types*, no one reached the highest level, and once the zero level was assigned. Additionally, six learners produced answers that were too short to be rated in this regard. As for *nominalisation*, four learners were rated on the lowest level, but six learners reached the highest level. Results for *choice of CDF types* are somewhat less problematic, with an average of 1.75 for group A and 1.95 for group B as well as six level-3 ratings and no zero-level ratings.

Zooming in on *choice of CDF types*, students frequently described or reported historical facts instead of making evaluations. For example, when asked to assess whether and how historical concepts<sup>39</sup> relate to the source provided, they often just reported something that they had probably heard in class. Sample 61 illustrates such a case:

61	RE: {Ancient Egypt was an advanced culture. They believed in many gods and where the first constructions of the irrigation systems. They had a central government which means that they had a king, in this case a pharaoh, which had the whole power and made decisions.}
Item 3	
ARJ01(A)	

In this example, the student summarized the corner stones of Ancient Egypt but failed to argue how the source could be connected to irrigation and central government. As he missed the main point of the task, his content ratings for these items were also very low with 0-1-3-1-1-2.

Especially results for *target level* but also *accuracy/relevance* and *target competence* seemed to be affected by employing presumably simpler CDF types. In terms of numbers, *choice of CDF type* correlated weakly with all content-related descriptors except for *systematicity* (weakest correlation with *justification/ comprehensibility*  $\tau_b = .30, p = .031$ ; strongest with *scope of content*,  $\tau_b = .43, p = .028$ ) when both data sets of the needs analysis are combined ( $N = 35$ ). Interestingly, when split, correlations were considerably stronger in group B, while in group A, *choice of CDF types* did not significantly correlate with any content domain.<sup>40</sup> This might stem from the smaller sample size of group A, where individual outliers have a greater impact. In combination with the qualitative analyses, however, the results presented above suggest that substituting CDF types could negatively impact content-related assessment, which is quite a logical conjecture from a conceptual point of view. By (mainly) employing CDF types that are associated with thinking skills on the reproduction level, such as DESCRIBE or REPORT, learners neglect higher-order thinking skills as defined in the history curriculum (i.e., reorganisation/ transfer and problem-solving/ reflection), which would often require them to EXPLAIN, EXPLORE, or EVALUATE.

Still, analysing only the appropriateness of the CDF type is not sufficient in terms of assessing content, as the following example demonstrates:

62	EV: {Ancient Egypt was part of the evolution of the human beings and all the technology because [RE: the Egyptians had the first paper, a calendar, first water system connected with the seasons,...]}
Item 4	
NNM05(A)	

In this example, NNM05 evaluates the historical role of Ancient Egypt and its impact on the “evolution of human beings”, when in fact she should have evaluated the validity of the source. In other words, while EVALUATE was indeed the target episode CDF type, she still missed the point of the task despite performing the intended episode CDF type.

Another indicator for content-related success seems to be the students’ ability to logically organise CDF types. In the needs analysis data of group B, this descriptor correlated moderately to strongly with content-related domains (weakest correlation with *target competence*  $\tau_b = .65, p$

<sup>39</sup> irrigation and central government in group A and colonialization in group B

<sup>40</sup> Overviews and exact values of all correlation tests can be found in the appendix repository (section II/ B/ folder [1 \(cycle 1\)](#), [2 \(cycle 2\)](#), and [4 \(needs analysis combined\)](#)).



= .003,  $n = 19$ ; strongest correlation with *justification/ comprehensibility*  $\tau_b = .75$ ,  $p = .001$ ,  $n = 19$ ). In group A, again, no statistically significant correlations could be detected in this regard.<sup>41</sup> Qualitatively, however, we see this relation also in data of group A, such as in example 63, which comprises a number of CDF types all sustaining the main episode CDF type:

63	EV: {I think it is true. [CA: [RE: Because slavery was common in ancient Egypt][DS: and the people on the picture look like slaves. The people also look like they are suppressed by that one person who has more power than them], [RE: which was the case in ancient Egypt, as well.]]}
Item 4	
ICM01(A)	

In this example, the student evaluates the validity of the source by comparing (CATEGORIZE) what she knows about the historical context (REPORT) to what she sees in the picture (DESCRIBE). As such, her basic CDF types all contribute to the overarching communicative intention of evaluating. Furthermore, the sequence and interlacing of the individual CDFs are comprehensible, as she also makes these connections explicit. On the content-scale, this answer received full points. In contrast, the following sample shows how an unclear CDF-composition contributes to an overall weak answer:

64	[RE: Since the moment in which Columbus travelled to India became the start for also other European Colonialists to continue the colonization. He was not the one to discover America as Amerigo Vespucci was there before him, [EV: but he was the first to make a significant change for the local people there (with colonializing). [EV: The point is that [EA/RE: he thought he reached India because his calculations were wrong.]]]
Item 3	
LED08(B)	

In example 64, identifying a main communicative intention, i.e., an encompassing CDF episode, is rather challenging. The student brought in various points somewhat related to colonialization but did not manage to arrange them logically. As a consequence, her evaluation of the source's relation to colonialization remains elusive. What she did evaluate in this example, however, was the role of Christopher Columbus. This could have been her strategy to assess the picture's connection to colonialization, but her rather confusing structure of CDF types combined with somewhat misleading linguistic markers does not really support this reading of her answer. For example, "the point is" would normally introduce some sort of evaluation, but in this case, she used it to REPORT another rather unrelated piece of information, namely why Columbus thought he reached India (EXPLAIN). In total, this answer received a rating of 1 across all content-related scales. Although LED08 provided quite a bit of information, she missed the central idea of the task and could not present this information in a way that would make clear what her main point really was.

As for *linking*, the students involved in this study received, on average, slightly better results in terms of *function* than in terms of *form*, surprisingly. Low *linking/form*-results mostly stem from absence or little use of linking in general paired with orthographic (e.g., "wherease" / SAA03), collocational (e.g., "but on the other side" / ARJ01) or syntactic mistakes (e.g., by having standalone subordinated clauses). As some examples presented above implied, good control of linking, especially from a functional perspective, contributes to the linguistic success of items. Appropriate

<sup>41</sup> In the aggregated data set ( $N = 35$ ), correlation scores between *composition of CDF types* and content-related descriptors are weak to moderate too (see appendix, [section II/ B/ folder 4 - needs analysis combined](#)).



linking signposts communicative intentions and sheds light on the relationship between different CDF realizations. An absence of linking, on the other hand, requires the reader to decode potentially implied communicative intentions and connections as illustrated by extract 65:

65	[RE: In Egypt the people had their own ruling system][DS: which quite shown in the illustration where the pharaoh rules and his people work for him.][EO: The workers might be from low class.] [RE: Egypts were far advanced in irrigation.]
Item 3	
ELH01(A)	

ELH01 basically just sequences different CDFs, and apart from the transition from REPORT to DESCRIBE in the first line, there are no markers pertaining to the relation of these different pieces. As such, no overarching communicative intention could be detected. Again, it can be assumed that she tried to EVALUATE the connection of the source to irrigation and central government implicitly, but there is no way of knowing for sure since she did not make these relations explicit.

Another issue repeatedly observed in this data was that learners tended to use “because” as a dummy linking device. There are many examples when learners introduced something they remembered from class, potentially hoping that using “because” would turn their statement into an explanation or justification. In other words, when “because” was used, learners often introduced reports or descriptions that were probably meant as reasons instead of actual reasons:

66	EO/EA: {I think he drew this [DS: because it shows how the relations and the roles were verteilt/=distributed/ in ancient Egypt.]}
Item 2	
NNM05(A)	

In this example, NNM05 described the picture rather than really providing or exploring motives for the production of the image and the stylistic choices of the artist. In general, learners tended to predominately use “because” and “so”, resulting in little variety of linking devices and inappropriate usage of these connectors. In terms of numbers, *linking in terms of function* and *form* correlated weakly with overall content outcomes ( $\tau_b = .35, p = .009, N = 35$ ;  $\tau_b = .36, p = .007, N = 35$ ). The correlations to individual descriptors were in a similar, rather weak range, with all correlations being statistically significant except for *accuracy/ relevance* and *systematicity*.

Turning to *hedging*, it was observed that a great number of learners simply did not hedge their claims at all, resulting in very low average ratings ( $M_A = 1.00, M_B = 1.11$ ). Out of 35 students, ten performances were rated as zero in terms of *hedging*, mostly because there was no evidence of it. Quite the contrary, some students seemed very sure of their statements and rather used emphasizing phrases than softening devices:

67	EV: {It a 1:1 portrait of how America was colonized. [RE: Italian Inquisitions explored America by accident and killed they own people of america and took advantage of them.]}
Item 3	
USN05(B)	

This student basically did the opposite of hedging by stating that the source is a realistic representation of the colonialization of the Americas, followed by a rather inaccurate and incoherent historical report.

Most performances (17 out of 35) were rated as level 1, meaning that they showed some evidence of hedging, but their use might be partly inappropriate. Sample 68 serves as an illustration:

68	RE/EO: {After they probably discovered an island they would have used the people that already lived on this island as slaves and would import them to America. [CA: Maybe this has also to do with the Dreieckshandel/triangular trade/, [DF: where the explorers start from Europe and import goods to Africa, so they can get slaves to take them to America and last but not least from America they would import cotton to Europe.]]}
Item 3	
ZE11(B)	

ZE11 overused and misapplied hedging devices. Combined with unclear personal deixis, her inappropriate use of conditionals and probabilistic adverbs obscured her communicative intention. Consequently, the reader cannot know whether she intended to hypothesize about the people depicted in the source or if she aimed at reporting something she knew about colonialization but was not entirely sure about its relation to the source. Put differently, her use of hedging devices did not serve its usual purpose in historical discourse, namely signifying that there are also alternatives to one's claim or interpretation.

As for correlations, *hedging* results correlated moderately with *overall content* results ( $\tau_b = .48, p < .001$ ) but with no other language-descriptors in the combined data set ( $N = 35$ ).<sup>42</sup> This means that performances with high content ratings tend to be characterised by a good control of hedging too. Interestingly, results for *hedging* appeared to be independent from other language-related descriptors.

Finally, compared to all other language results, the outcomes for *appropriate use of nominalisation* were not as problematic, with only four learners receiving a zero rating. Still, 15 out of 35 learners only reached level 1 of this scale, meaning that there was some evidence of nominalisation, but these might be used partly incorrectly. Here we have an example of such a performance:

69	DS: {The picture show's the arrival of Christopher Columbus and his crew in America and how they meet the native americans. Three christens in the back pull up a cross. On the sea are three ships and it seems that three people are running away. Also the "new arrivals" are getting presents form the native americans.}
Item 1	
ICK01(B)	

Task 1 asked the learners to describe the picture. ICK01 used one nominalised form in this answer, namely "arrival", while for the rest, she employed a more verbal style. This is very representative of the rest of her writing. Ten performances were assessed as level 2, and six texts reached level 3, meaning that for most learners, there was quite a bit of foundation to build on. An example for a level-3 rating is provided below:

70	DS: {I see Columbus, who has arrived in America. He is welcomed by the indigenous people of the land. The indigenous people bring gifts and very valuable goods. In the back, people are trying to put up a cross, [EO: which probably is related to the missionary-intentions of the European catholic church, [EV/RE: which was quite important]]. [DS: You can also see "Indians" running away from the ships,][EO: which could indicate the real reaction of the indigenous people.]}
Item 1	
UCQ07(B)	

<sup>42</sup> Apart from *systematicity*, all content-related descriptors correlated weakly with *hedging*. Again, please see appendix, [section II/ B/ 4 \(needs analysis combined\)](#) for exact numbers of the individual descriptors.

This student uses several nominalised forms, such as “missionary intentions”, “reaction”, as well as a gerund construction (“running”). As UCQ07 used a relatively nominal style throughout his writing, his use of nominalisation was rated as level 3.

Unlike all the rest of the descriptors, there were no significant correlations between *nominalisation* and other descriptors that were stronger than  $\tau_b = .30$ ,  $p = .048$  (*choice of CDF type*,  $N = 35$ ), neither in the combined nor the separated data set, meaning that the use of nominalisation did not relate to the level of sophistication of other aspects investigated in this study.

### **7.1.3.3 Summary and implications for the design**

The two groups under investigation yielded surprisingly similar results both in terms of overall scores as well as outcomes for specific descriptors despite their different experiences and attitudes towards BE. Both groups struggled more with the linguistic realization than with the performance of historical competences and thinking skills. Nonetheless, the data has shown that issues with the linguistic realization of their cognitive operations were indeed closely linked to their performance of historical skills.

Especially the ability to sequence and combine different CDF types in a way that supports their overarching communicative intention proved to be very challenging and, at the same time, vital for overall success. Although being less of a problem for these learners, employing appropriate CDF types in the first place also turned out to be important for a convincing demonstration of one’s subject-specific skills. Very often, when students were asked to perform higher-order historical thinking skills, such as reorganisation or reflection, the learners failed to perform appropriate CDF types, such as EVALUATE or EXPLORE. Instead, they tended to replace these with seemingly easier CDF types, like DESCRIBE or REPORT, which are more commonly associated with lower-order historical thinking skills, such as reproduction of declarative knowledge or recounts. In many cases, a failure to signal their communicative intention, i.e., the CDF type, contributed to low ratings because the purpose of the reported facts and/ or the description of visuals were not made clear. Thus, they could not build up convincing analyses or reflections, which detrimentally affected their content ratings. Low results for appropriateness of linking are closely related to this issue. The absence of linking devices appropriate for target CDF types often aggravated the problem of imprecise and implicit answers. Analysing written performances by business students in a mainstream and EMI setting, Breeze and Dafouz (2017) made a similar observation, stating that “the problem is precisely that the student fails to signal either CDF (DESCRIBE or EXPLAIN) explicitly, or to relate one to the other in a cognitively more complex operation” (p. 88). Thus, it could be argued that accurately signposting communicative intentions and their relations are issues relevant for different fields and educational levels, also from a subject-didactic perspective. Therefore, the intervention needs to focus on appropriate signalling of communicative intentions,

showing how one could convincingly work towards an overarching communicative intention (i.e., employing a logical composition of CDF types) and how different subordinate CDFs could be linked linguistically. Additionally, it seems paramount not to teach signalling phrases in isolation since the data also contained numerous examples of inappropriately applied signal phrases, tying in with Donato's (2016) call against pre-teaching of such phrases. Instead, the relation of 'useful' phrases to their underlying function needs to be highlighted and practised in meaningful subject-specific contexts.

One dimension where the above-mentioned issues come especially into effect is *justification/comprehensibility*. This descriptor, and by extension the whole concept of historical thinking, requires learners to produce a claim justified through its relation to historical evidence and knowledge about the historical context in a comprehensible and tangible way. To achieve this, one needs to demonstrate a range of different smaller cognitive operations, put them into a comprehensible linguistic shape, and connect them in a way that sustains their overarching communicative intention. Understandably, this is an extremely difficult task that almost all students struggled with. A study by Lorenzo (2017) examining written argumentative performances by 16-year old CLIL history learners illustrates this as well. He reported quite a number of "unsubstantiated opinions and without any real analysis or perhaps understanding of its purpose and effect" (2017, p. 37). This can also be seen in this data set since many learners struggled with drafting convincing and sound evaluations that are sustained by "real analysis" (Lorenzo 2017, p. 37), i.e., including and relating the source and its historical context, with extreme cases in which students only present their verdict without any form of (attempted) justification. Conceptually, this links to findings by SFL-based research, arguing that the proficiency in terms of interpersonal function, i.e., justified stance-taking, is essential for academic achievement, especially in higher grades (e.g., Christie & Derewianka, 2008; Schleppegrell, 2004). From a more subject-specific perspective, van Drie and van Boxtel (2008) report that learners of history tend to argue one-sidedly without considering counter-arguments and potential other interpretations. This relates to another problematic area for learners of history, according to van Drie and van Boxtel (2008) and Carretero and van Alphen (2014), namely that history students approach historical sources from today's perspective, ignoring their historical context. The data of this study and the literature presented here and in chapter 4 all point to the fact that engaging with historical sources in a subject-specific and adequate way is challenging. To support learners in such demanding tasks, one would need to unfold these connections, raise awareness, and support the learners' own production in a step-by-step manner with useful, contextually adequate linguistic tools and explicit instruction. Such scaffolding techniques could help learners unpack dense material and repack it when performing cognitively demanding operations such as producing a historically justified and convincing claim.

As for two features of historical discourse under investigation, hedging should definitely receive attention in the intervention since many learners avoided using hedging strategies or even presented their claims in a very absolute manner. Similar observations were also reported by

McCabe and Whittaker (2017) or Lorenzo (2017), where learners, too, struggled with hedging claims. This could negatively affect accuracy of content and might pose a problem for signalling their intent to EXPLORE or EVALUATE, which in turn could detrimentally influence their content results. Especially when working on CDF types that could benefit from good control of hedging, students should be given explicit instructions. Nominalisation, on the other hand, should not take up too much space within the intervention. The results for *nominalisation* do not warrant an in-depth treatment of nominalisations (1) because learner outcomes are relatively adequate and (2) control of nominalisation appears to be largely independent from the other dimensions under scrutiny. This seems to be somewhat in contrast to other studies, such as Nashaat Sobhy (2018), Morton (2010), or Lorenzo (2017), who all highlight the importance of nominalisations for subject-specific literacy skills. This discussion will be resumed in chapter 8 after the findings of the other research phases have been presented too.

## 7.2 Designing the intervention

In the following subchapter, the development of the two units designed in the course of this PhD study is outlined, including insights into how pedagogical decisions were made and realized in the materials designed, which aspects were considered, and how the intervention changed its form throughout the developmental phase of the study. Put differently, this part intends to illuminate the design phase, which is essentially an attempt at materializing and refining the insights gained by answering RQ1.

Unit I is concerned with the topic of absolutism and mercantilism and was developed in cycle 1 (context A) and cycle 2 (context B). Unit II deals with the Industrial Revolution (IR) and was created for context A (cycle 3). Before dealing with the didactic units individually, section 7.2.1 describes common threads and themes of all five<sup>43</sup> design sessions, shedding light onto collaborative pedagogical designing and planning. Here, central links to relevant DBR literature are provided to clarify the rationale of the design team. Then, the design process of each didactic unit is summarized in subsection 1 respectively, with the prime focus being on main decisions, which are exemplified with extracts from the materials. As unit I was implemented in two schools and thus went through two cycles, subsection 7.2.2.2 deals with the adaption for school B. The following subsections (7.2.2.3 and 7.2.3.2, respectively) then present the lesson plans to provide an overview of the design. Finally, the worksheets of the units are introduced and equipped with comments illuminating the didactic rationale behind certain decisions (7.2.2.4 and 7.2.3.3, respectively). To review the full design process, the reader is referred to the appendix repository ([section III/C - materials](#)). Here, all materials are available, including revisions of drafts and teacher's versions. Revised drafts show how the materials changed from initial draft to final version via the track-change function. Teacher's versions, on the other hand, offer a full key, tips for successful implementation, and justifications for pedagogical decisions. Lastly, [section II/C of](#)

---

<sup>43</sup> Unit I: Two meetings with T<sub>A</sub>, one meeting with T<sub>B2</sub>; Unit II: two meetings with T<sub>A</sub> = five design sessions

[the appendix \(design sessions\)](#) contains the summary table, MaxQDA code-matrices, hierarchical codes/sub-codes models, and code co-occurrence models of all design sessions.

### 7.2.1 Modus operandi of design sessions

Since the teachers involved in this project had limited time resources, the following procedure was arranged: First, we would collaboratively devise a rough outline for the unit and agree on a number of principles. These principles considered the participants' needs and wishes on the basis of preliminary results of previous data collections of the pilot and the ongoing main study. For example, the materials should be student-centred, interactive, cognitively engaging, and organised in small steps to counteract small attention spans and lack of motivation, making the process of 'linguaging' history more accessible to the learners. As in the pilot cycles, these processes should be unfolded and scaffolded on the basis of CDFs, including language support in the form of glossaries, useful phrases for the CDFs in question, and explicit advice on how to implement these phrases, what certain language functions entail, or on historical discourse more generally. All these so-called *language boxes* target difficulties as identified in the needs analysis, such as how to signal communicative intentions or hedge a claim when evaluating a historical source. In line with these agreements, I drafted a lesson plan and materials, which I then presented to the respective teacher, who provided feedback and offered suggestion for improvement, which in turn were the basis for the revision of the drafts.

In all design sessions, the teachers offered their expertise, aiming at increasing feasibility and educational value, which are central aspects of designing in a DBR setting, according to Euler (2014). Regarding feasibility, the teachers offered suggestions on how to increase clarity of prompts and potential ways of individualisation to cater for different learning paces. They also commented on the number of exercises, appropriateness of difficulty of content and language, as well as inclusion or exclusion of sub-topics and the respective exercises. Concerning pedagogical value, the teachers evaluated the overall approach as well as the individual tasks. Both teachers felt that CDF-based scaffolding might be a way to raise awareness about the role of language for content subjects and to support the learners' skills development. They nonetheless offered suggestions concerning the improvement of the tasks in terms of pedagogical value, usually with the aim of achieving learning goals and ensuring working knowledge, for example by maximising clarity of tasks, avoiding overload or redundancy, or by setting the right expectations, as illustrated in the extract below:

71	English translation	Original quote
T <sub>B2</sub> Session 3	And if you make it clear to them that it [=our approach] is about content and how you approach sources, then I find these boxes extremely valuable.	<i>Und wenn man ihnen klarmacht, dass es um zwei Sachen geht, dass es um den Inhalt geht, aber auch um eine Herangehensweise an Quellen geht. Dann finde ich diese Kästchen irrsinnig wertvoll.</i>

Here, T<sub>B2</sub> appreciated the language boxes featured in the materials but only if their purpose was made clear to the learners by communicating these aspects openly. When offering suggestions or evaluations of tasks, the teachers also seemed to have their learners' usual behaviours and skill profiles in mind:

72	English translation	Original quote
T <sub>A</sub> Session 5	<p>TA: Please add “formal” [...] because this is something I can’t [say] often enough.</p> <p>R: Tentative, polite ((reads prompt))</p> <p>TA: Mhm [affirmative]. They won’t understand “tentative”.</p> <p>R: It was in the first unit, where it explains</p> <p>TA: =((laughs)) [...] No, they won’t know it anymore.</p>	<p>TA: <i>Schreibst bitte dazu „formal“ [...] weil das ist etwas, das kann ich auch nicht oft genug.</i></p> <p>R: <i>Tentative, polite. ((liest Angabe))</i></p> <p>TA: <i>Mhm [zustimmend]. „Tentative“ werden’s nicht verstehen.</i></p> <p>R: <i>Es ist aber im vorigen schon einmal vorgekommen, da erklär</i></p> <p>TA: <i>=((Lacht)) [...] Nein, das wissen sie nicht mehr.</i></p>

In this example, T<sub>A</sub> intended to clarify a prompt based on her knowledge of the students’ language skills, namely that they tended to use informal language in formal contexts, which also corresponds to what the students said in a previous interview. She further implied that explaining a term once usually would not suffice, in this case “tentative”, and would therefore require clarification again.

Euler (2014) mentions that teachers and researchers should discuss assumed user competences required for the implementation of the design, potential critical events, and their solutions, as well as the possibility of adjustments under changeable conditions. These are all points which were addressed by the two teachers. For example, it was decided to create detailed teacher’s versions with comments and explanations of pedagogical decisions to ensure that other teachers could also use these materials and adapt them for their purposes. Another thought brought up by both teachers was to debrief between lessons to ensure smooth implementation, allow for adaptations if necessary, or collect ideas for the commented version of the materials. These debriefings also offered the opportunity to collate interpretations of what happened in the classroom (Gravemeijer & Cobb, 2006) and ensured that the interests of practice and research did not diverge (Euler, 2014). To make sure that interests would not deviate throughout the course of the project, I sometimes brought up the overall rationale and scope of the research project or previous preliminary results during the design sessions, which the teachers seemed to welcome and sometimes even initiated. Conversely, the teachers’ practice-related concerns and wishes were considered as much as possible, ensuring that the project would not interfere with their annual programmes and normal schedules.

On the whole, all design sessions were productive and cooperative, often following the same patterns of introducing a task or source, a discussion thereof, including reflections and a number of suggestions, elaborations, and sometimes counter-suggestions, but always resulting in agreement. Apart from these evaluations, decision-making processes and reflections, the design sessions were used for organising the implementation phase as well as other research steps, e.g., the post-intervention tasks, such as booking computer rooms or printing copies. At the end of each session, upcoming tasks were divided and deadlines specified. Usually, it was my task to further revise the materials based on our discussion and consult other researchers outside the project (as recommended by van den Akker & Nieveen, 2016) while the teachers provided feedback once more via our online file-share repository and made sure that everything was in place for the

implementation of the materials. To organize this process, we stayed in contact via telephone and e-mail.

## 7.2.2 Unit I: absolutism and mercantilism (cycle 1 & 2)

### 7.2.2.1 Design process of unit I for school A (cycle 1)

For unit I, we agreed to create a unit on the topic of absolutism and mercantilism. This topic reflects the Austrian history curriculum for this school type<sup>44</sup> as well as the teacher's annual programme and could be realistically covered in four to five lessons. Since this study focuses on methodological competences and orientation competence, source analyses are at the core of this unit, which also overlaps with the curriculum: "The students can use **historical sources** critically to **reconstruct** and **deconstruct** history" (Austrian Federal Ministry for Education, 2014, p. 91-92).<sup>45</sup> To be able to do this, the learners might need to DESCRIBE, EXPLAIN, CATEGORIZE, EXPLORE, REPORT, or EVALUATE (see Bauer-Marschallinger, 2016). Consequently, it was decided to let the learners work on historical sources dealing with Louis XIV and absolutism and to scaffold their de- and reconstruction on the basis of the CDFs needed for the learning objectives specified in the curriculum in small steps, similar to what was done in the pilot cycle. Figure 14 provides an illustration of our approach. In this example, the learners were provided with a memoir written by Louis XIV aimed at his son. The task was to argue whether Louis' description of an 'ideal' king could have corresponded to himself. As already outlined in section 7.1.3, learners seemed to struggle with providing reasons for their views. To address this issue, this task is accompanied by an awareness-raising remark on the importance of justification for historical argumentation as well as by practical phrases for expressing views and providing reasons. As such, this task targets the CDF type EVALUATE.

3. **Argue** whether the 'real' Louis XIV could have corresponded to his description of an 'ideal' king (as he outlined in paragraph 2). Don't forget to **provide reasons** for your judgement!

Take notes (keywords) and once you talk to somebody, try to use some of the phrases from the box below:

expressing views	providing reasons
in my view	because (of)
to my mind	since/ as
my impression is that	due to
it seems to me that	owing to (the fact that)...

In history it's very important to **clearly state** if something is **your opinion** rather than an established fact.

Yet, history is not really about opinions, but about **justified** views! It's very important to **base your views on reasons**.




Figure 14. Worksheet Louis XIV, source A, task 3

<sup>44</sup> See *Austrian Federal Ministry for Education* (2014, p. 92): "**Fundamentals of the modern state** and **approaches of implementation** as well as crosscurrents (antique models, bourgeois revolution and restoration, **forms of rule and leadership structures**), nation building".

<sup>45</sup> All quotations of the curriculum are taken from its official translation (Austrian Federal Ministry for Education, 2014). Emphasis in **bold**, however, was added by the author of this thesis.



Apart from focusing on CDFs, the intervention would also provide linguistic support on linking devices and two features of historical discourse relevant for some of the CDF types. These two features are nominalisations to increase concision, formality, and level of abstractness of the students' output (especially central for REPORT and EXPLAIN) and hedging to make their claims less absolute and more qualified (particularly important for EVALUATE and EXPLORE; see also Coffin, 2006, and section 3.2.2).

In order to equip the learners with the necessary working knowledge, we decided that we should include an introductory phase in which the learners gain a rough conceptual understanding of absolutism. To reactivate previous knowledge, T<sub>A</sub> suggested to start with a quick brainstorming activity on the term "absolutism", which would serve as a basis for a subsequent whole-class discussion on past and current examples as well as key aspects of absolutism, supported by some slides. We also contemplated to include a summarizing reading on topics like the Thirty Years War or Reformation as well, which could be useful to understand the preconditions of absolutism, but we decided against their inclusion since the curriculum specified them for later semesters and we wanted to keep the introductory phase as short as possible to leave enough time for the scaffolded source analyses. We further considered the possibility that the students might not contribute much during the introductory input phase. In this case, the teacher would just present these contents in the form of a lecture. In any event, we also determined that the teacher should not provide too much information on Louis XIV and his life at Versailles since we would rather have the learners work on this topic with a number of different sources, providing different perspectives on this archetype of absolutistic rulers. This would involve "**describ[ing]** different forms of rule and leadership structures and **discuss[ing]** their effects on states and the society", as specified in the curriculum (Austrian Federal Ministry for Education, 2014, p. 92).

A preselection of sources was based on materials I had designed when I taught CLIL history earlier, but the actual choice, nonetheless, took quite some time. Initially, the teacher suggested to let different learner groups work on different sources to be able to include a variety of perspectives, but eventually, we decided against this group format based on our experiences in the pilot cycle. As mentioned in chapter 6, the pilot cycle was organized as one big group jigsaw activity, in which each group worked on one topic in a scaffolded way and then shared their results with the rest of the class. As a result, students only felt knowledgeable in 'their' topic and were overwhelmed with their peers' presentations. The reason for this could be that the input presented by their colleagues was not scaffolded and no active engagement was required, resulting in miscomprehension of historical concepts. Another reason why we abandoned the idea of splitting them into groups and letting them work on topics independently was that we got the impression that the learners skipped the more difficult parts and did not consider the language tips. Since everybody completed different tasks with slightly different linguistic input, the teacher could not systemically check these aspects. Consequently, it was decided that this time, all learners would work on the same sources since this format reduces the risk of learners skipping tasks and allows a systematic comparison of results and reprocessing of important aspects, including the

language tips. In the end, we settled on the type of sources to include, namely a self-descriptive text by Louis, one portrait of him, and one or two texts presenting Louis from an outside perspective. We also agreed that the analyses could be done in pair work and should, at some point, be compared in a plenum setting.

After devising this rough outline of the first part of the unit, we moved on to the topic of mercantilism. Here, T<sub>A</sub> suggested that we could work with a flowchart explaining the mechanisms of a mercantilist economy. This would correspond to another can-do statement of the curriculum, namely: “The students can [...] **describe** and **compare** ideal models and concrete economic systems by means of their characteristics” (Austrian Federal Ministry for Education, 2014, p. 92). The teacher reported that in the past, she had worked with a flowchart by summarizing the chart in the form of whole-class teaching but had been disappointed when testing this topic:

73	English translation	Original quote
T <sub>A</sub> Session 1	Yes, both groups last year really struggled with this. Because for me, it's somehow totally logical and I thought 'a flowchart, excuse me describing a flowchart' [is really easy] but yes, it was not that easy.	<i>=Ja, da haben sie sich voll schwer getan letztes Jahr, beide Klassen. Weil für mich war das irgendwie total logisch und ich hab mir gedacht ein Flowchart, entschuldige, ein Flowchart beschreiben aber ja aber das war nicht so easy.</i>

The teacher added that the learners would require quite a bit of linguistic know-how to be able to verbalize the processes depicted in the graph. With this in mind, we decided to include the flowchart in our design with the hope that CDF-based scaffolding could provide relief in this regard. The teacher and I also thought about whether we should implement the linguistic support before actually interpreting the flowchart or whether it made more sense to integrate it into the task. Initially, the teacher suggested going through the linguistic support as preparation for the analysis of the chart because this would ensure that all students had language-focused instruction before starting the task. At this point in time, this seemed plausible but when going through the insights of the pilot cycle once again after this design session, this part was rearranged. One insight of the pilot study was that the linguistic support should be tailored to the task at hand, integrated into the materials, and not too comprehensive. In the pilot study, especially weaker students seemed to struggle with large chunks of linguistic help and deciding which aspects were relevant for which part of the task. Therefore, it was later decided to integrate the linguistic support into the materials and to make sure that the language boxes were purposeful for the task at hand, which also is a decision consistent with T<sub>B2</sub>'s point to cater for short attention spans and with Donato's (2016) call against pre-teaching relevant language. Apart from the rough outline of the unit, we also agreed that it would make sense for the teacher to really monitor the students' use of language production and comment on it if learners struggled with any of the linguistic features targeted by our intervention. This might involve asking for full sentences, more precision, or recasts using another phrase.

Typically, the first design session concluded with discussing next steps and clarifying responsibilities. Since many history lessons had to be cancelled in this group, with only one history lesson per week in the winter term, we had to postpone the first intervention to the second

semester, in which two history lessons were scheduled per week. Thus, the second design session with T<sub>A</sub> took place as late as early February.

In the second session, I presented my first draft to the teacher and outlined my didactic decisions that influenced my design, seeking the teacher's feedback in order to create a feasible and ecologically valid design with clear and precise prompts. Therefore, a substantial part of this session was dedicated to going through the tasks and specifying and reformulating their prompts. This included clarifying where we would like the learners to produce full sentences and where key words would be more appropriate. We agreed that key words should be used with tasks more prone to oral language, especially those that could benefit from interaction. Full sentences should be used for tasks which were cognitively challenging, requiring sophisticated language, and/ or focused on a CDF type the learners needed more practice with. Another aspect we considered was ensuring variation between key words/ oral tasks and full sentences/ written tasks. In the course of this discussion, we also finalized our planning in terms of social format, i.e., which tasks should be done alone, in pairs, or in groups. Figure 15 below provides an example of how we specified social formats and whether students should produce key words or full sentences:

a) Paragraph 1: ~~Discuss~~ Describe the work ethic/ routine of Louis XIV, according to himself. Take notes and discuss with a partner. ~~How does Louis XIV describe his own work routine/ ethic?~~

⇨ b) Paragraph 2: Together with a partner, underline all the feature of a good king and all the things he should do, according to Louis XIV. Then, in your own words, try to outline your findings in full sentences, using some of the words from the box below.  
~~Then try to outline your findings using some of the words in the boxes.~~

<b>organizing ideas</b>	<b>adding information</b>	<b>summarizing</b>
first of all	moreover	to sum up
firstly/ secondly/ ...	furthermore	generally speaking
finally	additionally	in summary
closely connected to	another central feature/ idea/ ...	in brief

**Why?** Using linking phrases helps organize your thoughts. As a result, others can easily follow your summary!

**How?** Start with the most important, and then gradually add other ideas (and mark them as other ideas by using these phrases). Finally, mark your concluding statement by using summarizing phrases.

Figure 15. Revision: worksheet Louis XIV, source A, task a and b (later relabelled as task 1 and 2)

Figure 15 also provides an example for another facet that needed revision, namely the use of performative verbs for the prompts. At the suggestion of the teacher, we transformed wh-question to prompts containing performative verbs, e.g., replacing the question previously used for task 1 above with DESCRIBE. The teacher explained that

74	English translation	Original quote
T <sub>A</sub> Session 2	[t]he reason why I'm so picky [concerning the use of performative verbs] is that I'm really working [on this topic] with them and it's a really tough process for them to get it [...] I've explained it in theory (.) but yes, that's why I would like it to be consistent.	<i>[w]arum ich da so picky bin, weil [...] ich gerade voll damit, also mit ihnen arbeit, äh, und das ist so ein zacher Prozess, dass sie das checken [...] ich hab's in Theorie durchgemacht (.) aber, ja und deswegen hätte ich gerne, dass das einheitlich ist.</i>

In this design session, T<sub>A</sub> mentioned once more that her learners really struggled with responding appropriately to tasks containing performative verbs although she had explained them in theory again and again, without satisfactory results, which is an observation also backed up by the students in their interviews. Consequently, T<sub>A</sub> wanted to incorporate performative verbs consistently in the task design to better prepare them for testing situations. Additionally, by changing the wording and discussing which performative verb was most suitable, we could make sure to highlight the underlying communicative intention we had in mind. We also talked about whether the prepared language tips were appropriate for the target group and the tasks they accompanied and to what extent these boxes were different compared to the pilot materials. Since the students of the pilot study felt that the tips were somewhat irrelevant or difficult to realize, some more input was provided, highlighting why these linguistic tools are important for the language of history or how they can be applied in this context (see Figure 15, task b, grey box). The teacher welcomed these additions since she felt that learners would not be aware of these aspects of historical discourse without explicit instruction.

In the second design session, we also discussed the role of vocabulary and how it should be dealt with. Since the students of the pilot study appreciated the glossaries used in the pilot materials, potentially unfamiliar vocabulary was again clarified in footnotes. We further agreed that students were allowed to use online dictionaries, but no activities focusing on vocabulary should be included since this was not the goal of this unit. Furthermore, the teacher approved of the words to be included in the glossary. We also discussed the appropriateness of the texts to be included for the target group, with the result that some parts could be challenging but, with the help of the glossary, the scaffolding, and tips, manageable for these learners. To ensure feasibility, we cut parts of source B since we felt this source was too long and would require another task if we kept its original length and we agreed that we had enough activities and tasks already. We also talked about time planning and agreed on time frames for each source and work phase that seemed realistic to us based on previous experiences and knowledge of the students. Nonetheless, we were aware that our time frames might need flexible adaption when teaching this unit.

We also made decisions in terms of organising the workflow and supervising the learning process. In this respect, it was agreed to do a whole-class comparison after each source, in contrast to after each task or after the whole worksheet was completed. The reason for this decision was to monitor the students' progress and make sure they were focused on the tasks at hand without disrupting their workflow too much. Finally, we agreed on a procedure to finalize the unit and the materials. It was decided that, like last time, I should work in all the decisions of the second design session but also seek feedback from other researchers outside the project and incorporate these with track-changes. These materials were then made available to T<sub>A</sub>, who reviewed these modifications. Most of these changes were concerned with layout or minor reformulations, but it was also suggested to add concrete examples to some of the language boxes, which the teacher and I found to be a valuable addition (see Figure 16 for a sample).

e) Look at the last two lines of paragraph 2. ~~Why do you think does~~ Discuss possible reasons why Duke Saint-Simon ~~mention~~ mentions that Louis' ministers were "drawn from the non-noble class"?  
 Write down your answer in a full sentence using a strategy from the box onto the right.

Whenever we talk about something **hypothetical**, i.e. something we can't know, we need to show explicitly that we are **speculating**. We can do this by using modal verbs (**could, would, may, might**). We could use also words expressing probability, such as **probably, maybe** or **likely**, or construct **either-or sentences** or **if-sentences**.  
 E.g.: Considering what we know about Marie-Antoinette, she really might have said her famous quote "let them eat cake". Yet, there is no evidence, so she probably never said it.

Figure 16. Revision: worksheet Louis XIV, source B, task b (later relabelled as task 2)

### 7.2.2.2 Adaption of unit I for school B (cycle 2)

I met with T<sub>B2</sub> after the implementation of unit I in school A. This design session was intended to improve the first version based on preliminary results and experiences from cycle 1 as well as to adapt the design to the needs of group B. On the whole, T<sub>B2</sub> found the materials effective and well done but also challenging and relatively comprehensive:

75	English translation	Original quote
T <sub>B2</sub> Session 3	I think [...] it's quite laborious for the students because it requires high levels of concentration. I wouldn't really change anything for now. I appreciate the sequence, the work steps. [...] Your task with these materials is to really change something in the learners and that's what's exciting.	<i>Ich find's [...] sehr anstrengend für die Schüler, weil sie doch immer hohe Konzentration verlangt. Ähm, ich würd jetzt auf die Schnelle eigentlich nicht unbedingt was ändern dran. Also ich find, die die Abfolge gut, ich find die Arbeitsschritte gut. [...] deine Aufgabe ist ja, dass, dass du Materialien hast, die ja dann wirklich auch was verändern mit den Schülern und Schülerinnen und das ist ja das spannende dran.</i>

His assessment of the materials as laborious and challenging but, to some extent, also as transformative corresponds to what was observed in cycle 1. In the first cycle, the learners' performance of historical skills and academic literacy indeed improved to some degree (see section 7.1.3), but some students struggled with the materials and needed more time than anticipated. Consequently, the students' paces differed considerably so that, overall, more time was needed for the individual parts of the worksheet on Louis XIV. In the end, source D was left out altogether in cycle 1 in order to be able to keep to the schedule. To account for varying speeds, T<sub>B2</sub> suggested a different organisation of the tasks on Louis XIV. Instead of having all students complete all tasks, source D (which is also somewhat more focused on historical literacy and thus new to them) should be optional, catering to quick learners. For this to work, we added the following instruction:

There are four different sources in connection to Louis XIV. Have a look at them and do the tasks below.  
 Procedure: First, do the tasks on your own. Once you've finished all tasks related to the source in question, raise your arm to find a partner to compare and discuss your results.

Figure 17. Revised instructions: worksheet Louis XIV

This way, learners can work on the tasks in their own time and exchange their ideas with

different peers, not just their neighbours. The teacher also added that the learners should be instructed to note down with whom they worked with to stress each individual's accountability.

This reflects one issue observed in the other group. In the students' retrospective interview, the learners said that the instruction "take notes and talk to your partner" did not make them feel responsible for their own output, resulting in very superficial notes.

In this design session, we also talked about the implementation phase and which steps would be important for a successful unit. Again, experiences with group A were taken as a starting point for this discussion. In group A, the learners were somewhat overwhelmed with the labour-intensity of the unit. Relating to this, T<sub>B2</sub> mentioned that his students were not used to writing and actively working for longer periods in history education but were, on the whole, capable of sustaining longer working phases. He further argued that since the materials comprised different methods, these learners should be able to stay focused, provided that one mentally prepared them for the upcoming challenge. T<sub>B2</sub> rightfully argued that setting the right expectations, including highlighting the purpose of the following work-intensive activities, at the outset of such a unit is quite important to avoid frustration. By stressing the purpose, the students might be more motivated to put in the required effort. T<sub>B2</sub> added that a 'pep talk', explaining the relevance of language for the subject history, could steer the learners' attention to these features and increase the probability of them considering these boxes when working on the tasks. Previous cycles had shown that learners tended to ignore language support and extra information if their purpose was not made explicit. To avoid adding yet another input box or elaborating the initial instructions, which some learners might overlook, we agreed that the communication of purpose should be done orally by the teacher. Therefore, remarks pertaining to this issue were included in the teacher's version of the materials (see subsection 7.2.2.4 and [appendix section III](#)).

Concerning the implementation, we discussed the use of German and the extent of code-switching. The teacher seemed somewhat concerned that his English skills might not suffice and thus announced that he would switch to German regularly. In general, the teacher wondered about how prominent his role should be, considering that the more he supported learners when working with the materials, the less the results spoke for the materials alone. In these respects, we agreed that the implementation phase should be realistic and tailored to the participants' needs. In other words, as the intervention should be ecologically valid, the teacher should use these materials like he normally would, bearing in mind his expertise and experience with the learners of this group. These practical insights can then be included in the teacher's version of the materials to help other teachers with the implementation of the activities. To increase feasibility of the unit, T<sub>B2</sub> and I agreed to cut two activities, namely the brainstorming in the beginning, which would require quite a bit of time if done properly, and exercise C/3 (comparison of sources, see appendix, [section III/1/file 5 – unit 1 all revisions](#)), which did not lead to very substantial results in group A. Task 6 of the mercantilism worksheet was modified, including the omission of the term "nominalisation" to make it more accessible to the learners.

Finally, we also discussed practicalities, such as scheduling or the use of laptops. We decided to let the students work on their laptops rather than with paper and pencil simply because the learners of group B were more used to working digitally and thus usually preferred it.

### 7.2.2.3 Final lesson plan of unit I

grade	number of lessons	language
11/ 3 HAK (secondary colleges of business administration)	4-5	English (German)

#### Themes:

→ curriculum 5<sup>th</sup> semester: “**Fundamentals of the modern state** and **approaches of implementation** as well as crosscurrents (antique models, bourgeois revolution and restoration, **forms of rule and leadership structures**), nation building” (Austrian Federal Ministry for Education, 2014, p. 92)

**Topics:** French absolutism & mercantilism

→ focus on Louis XIV, life at court, what it means to rule absolutely, mercantilism

#### Objectives:

Curriculum, 5<sup>th</sup> and 6<sup>th</sup> semester (Austrian Federal Ministry for Education, 2014, pp. 91 -92):

*The students can*

- *use historical sources critically to **reconstruct** and **deconstruct** history* (DESCRIBE, EXPLAIN, CATEGORIZE, EXPLORE, REPORT, EVALUATE)
- ***describe** significant historical processes of change, **analyze** and **explain** their causes* (DESCRIBE, EXPLAIN, EXPLORE, REPORT, EVALUATE)
- ***describe** and **compare** ideal models and concrete economic systems by means of their characteristics* (DESCRIBE, CATEGORIZE)

Further, discipline-specific language focus (see also subsection 3.2.2.2):

The students can

- use **nominalisations** appropriately to increase concision, formality, and level of abstractness of their output
- **hedge** their ideas to make their claims less absolute and more qualified
- **link** their ideas effectively (appropriate choices and form)

rough time frame	procedure/ content	interaction format	skills / language system	language	historical competence	materials	learning phase
5'	Quick (re-)introduction of the term “absolutism”	plenum	speaking vocabulary	DEFINE (DF), REPORT (RE)	FC (factual competence)	slides (absolutism)	introduction
25'	Presentation and discussion of key elements of absolutism	plenum	listening speaking	DF, CATEGORIZE (CA)	FC	slides (absolutism)	input
100'	Scaffolded analysis of <b>source A (memoirs by Louis XIV)</b> . Task 1: describing the routine of a king	students work individually- after each source, they look for a partner to compare and discuss results	speaking reading writing	DESCRIBE (DS)	MC (methodological competence)  (FC)	WS Louis XIV	elaboration
	Task 2: virtues of a king			RE, DS, linking & structuring summaries			
	Task 3: comparing ideal descriptions of a ruler to real historical figures (Louis XIV)			EVALUATE (EV), CA, DS, RE, linking for comparisons			
	Scaffolded analysis of <b>source B (Memoirs by Duke Saint-Simon, noble living at court)</b> Task 1: identifying criticism, reporting views		speaking reading writing	RE, EXPLAIN (EA), linking for cause-and-effect	MC, (FC)	WS Louis XIV	elaboration
	Task 2: analysis of motives, close-reading			EA, EXPLORE (EO), hedging			
	Scaffolded analysis of <b>source C (painting of Louis XIV)</b> Task 1: describing the portrait		speaking reading writing	DS, CA	MC	WS Louis XIV	elaboration
	Task 2: analysis of motives, purpose of painting			EA, EO, hedging			
	Scaffolded analysis of <b>source D (text by historian);</b> only for the fast ones Task 1: identifying nominalisation in historical writing		reading	nominalisation	MC (OC) (orientation competence)	WS Louis XIV	elaboration
	Task 2: reporting different views in historical writing			RE, reporting verbs			
	Task 3: different ways of reporting views in historical writing		writing				



	Task 4: identifying a historian’s evaluation of a historical figure		speaking writing	EV, RE, EA			
20’	Final task: <b>assessing the significance of a historical figure</b>	individual, then plenum	speaking writing	EV, CA, EA, RE	OC	WS Louis XIV	summarizing task
40’	Working with a <b>flowchart</b> on mercantilism Task 1: definition of mercantilism	pair, then plenum	speaking writing	DF	FC	WS mercantilism  slides (mercantilism)	elaboration
	Task 2: explanation of increased production	pair, then plenum		EA	MC, FC		
	Task 3: consequences of protective tariffs	pair, then plenum		EA			
	Task 4: explanation of cheap production	pair, then plenum		EA, RE, ordering factors			
	Task 5: reporting positive outcomes	pair, then plenum		EA			
	Task 6: recap flowchart/ mercantilism	plenum	EA, RE, DF	summarizing			
	Task 7: discussing disadvantages of mercantilism	pair, then plenum	speaking		EV, EA, EO	MC, OC	
10’	Recap: <b>revision of central insights and discussion of historical implication</b> (→How is this important to us/ connected to today?)	plenum	speaking	EV, RE, EA	OC	Blackboard/ <i>Padlet</i> (app)	consolidation

#### 7.2.2.4 Final version of worksheets of unit I

Two worksheets (WS) are presented in the following: worksheet 1 on Louis XIV of France and worksheet 2 on mercantilism. Empty space for students' answers has been reduced to present the materials in a more condensed layout. Furthermore, comments are provided in grey boxes, explaining some of the decisions that shaped the design. In the teacher's versions, available in the appendix file ([Section III/A/3 – key and annotations](#)), these justifications of pedagogical decisions are reflected in the annotated version of the materials, alongside a key and various tips for a successful implementation.

## LOUIS XIV OF FRANCE

There are four different sources in connection to Louis XIV. Have a look at them and do the tasks below.  
*Procedure: First, do the tasks on your own. Once you've finished all tasks related to the source in question, raise your hand to find a partner to compare and discuss your results.*

**Comment:** Not everybody has to do all four sources. One should set a time frame and let them work at their own pace. Once they are done with one set of tasks, they can look for somebody with similar speed to compare and discuss results (or to do those tasks that should be done in pairs).

This individualisation strategy should make sure that learners can work at their own pace and that they exchange results with different peers, not just their usual partners. This way, they might engage with a variety of learners, possibly with different levels of ability but same speed.

Initially, the learners might need a bit of support in arranging these processes, i.e., when looking for partners. Potentially, the teacher needs to intervene if they still continue to only talk to their usual partners.

**A) Memoirs:** *Description of Kingship*<sup>46</sup> written by Louis XIV for his son, describing the functions and behaviour of a king.

- 1 Without any doubt, two things were absolutely necessary for ruling: very hard work on my part, and a wise choice of persons who were capable of carrying out my work. I set a rule for myself to work regularly twice each day for two or three hours at a time. Each time I worked with different persons. This regular work did not include the hours which I spent privately working on matters of state, or the time I was able to give on particular occasions when special problems arose and I permitted people to talk to me about urgent problems at any time. My *timidity*<sup>47</sup>, especially on occasions when I had to speak in public, disappeared in no time. I felt that I was king and born to be one. I experienced a delicious feeling which you will not know until you are king.
- 2 A king must be guided by his own good sense, which is natural and effortless. [...] There is no satisfaction equal to that of noting every day some progress you have made in glorious and *lofty*<sup>48</sup> enterprises and in the happiness of your people which comes from the work you have done yourself. My son, the work of a king is agreeable. One must have his eyes open to the whole earth. He must *endeavour*<sup>49</sup> to learn each hour the news concerning every province and every nation, the secrets of every court, the moods and weaknesses of every prince and every foreign minister. He must be well informed on all matters from *commerce*<sup>50</sup> and science to art and philosophy. He must find out the secrets of his subjects, and discover the selfish interests of those who approach him with their real motives disguised. I know of no other pleasure I would take in place of the work of a king.

**Comment:** Words presumed to be unfamiliar to the target learners are marked in italics and translated in footnotes to ensure that the students can follow the text. Since the language learning goals of this unit are not focused on vocabulary, translations are enough.

---

<sup>46</sup> Available online:

[http://www.northernhighlands.org/cms/lib5/NJ01000179/Centricity/Domain/58/louis\\_XIV\\_Primary%20source\\_s.pdf](http://www.northernhighlands.org/cms/lib5/NJ01000179/Centricity/Domain/58/louis_XIV_Primary%20source_s.pdf) (please note that in the original worksheets, the footnotes start at 1)

<sup>47</sup> Zaghftigkeit, Schüchternheit

<sup>48</sup> erhaben, vornehm

<sup>49</sup> bestreben

<sup>50</sup> Handel

## TASKS A

First, do the three tasks on your own. Once you've finished task 1-3, raise your hand to find a partner to compare and discuss your results.

1. Paragraph 1: **Describe** the work ethic/ routine of Louis XIV according to himself. Take notes.

**Comment:** Bold type is used to guide the learner's attention. In this case, it is supposed to highlight the type of language function that is expected of the students. This should raise their awareness and, in some cases, help them connect the performative verb to the language support box(es).

2. Paragraph 2: Features of a good king:

- First, **underline** all features of a good king and all the things he should do according to Louis XIV.
- Then, try to **outline** your findings in full sentences, using some of the words from the box below.

organizing ideas	adding information	summarizing
first of all	moreover	to sum up
firstly/ secondly/ ...	furthermore	generally speaking
finally	additionally	in summary
closely connected to	another central feature/ idea/ ...	in brief

**Why?** Using linking phrases helps organize your thoughts. As a result, others can easily follow your summary!

**How?** Start with the most important, and then gradually add other ideas (and mark them as other ideas by using these phrases). Finally, mark your concluding statement by using summarizing phrases.

**Comment:** The coloured boxes provide awareness-raising remarks to communicate to the learners how the phrases in the box are relevant for historical literacy. Furthermore, these boxes sometimes also provide some more explicit instruction how to actually use the phrases in the box.

The boxes with phrases are intended to facilitate the learners' production when trying to perform the target language function. To avoid overwhelming learners, these should be tailored to the task and not too extensive.

In general, the boxes with linguistic support represent what learners seem to struggle with according to the needs analysis.

3. **Argue** whether the 'real' Louis XIV could have corresponded to his description of an 'ideal' king (as he outlined in paragraph 2). Don't forget to **provide reasons** for your judgement!

Take notes (keywords) and once you talk to somebody, try to use some of the phrases from the box below:

expressing views	providing reasons
in my view	because (of)
to my mind	since/ as
my impression is that	due to
it seems to me that	owing to (the fact that)...

In history, it's very important to **clearly state** if something is **your opinion** rather than an established fact.

Yet, history is not really about opinions but about **justified** views! It's very important to **base your views on reasons**.

**\*\* Now raise your hand to find a partner to discuss and compare your results! \*\***

For task 3, if your partner mostly disagrees with your view, write down his or her arguments too.

**B) Memoirs** by the Duke Saint-Simon, a French noble living at the court of Louis XIV<sup>51</sup>.

- 1 Louis XIV made for a brilliant court. His figure, his grace, his beauty, his grand *bearing*<sup>52</sup>, even the tone of his voice and his majestic and natural charm set him apart from other men as the king. Even if he had been born a simple private gentleman, he still would have excelled in all social festivities.
- 2 However, intrigues against the king during his childhood made Louis suspicious of intelligent, educated, noble, and highly *principled*<sup>53</sup> men, and as he advanced in years, he began to hate them. He wished to reign by himself, and his jealousy on this point soon became a weakness. The superior ability of his early ministers and generals soon *wearied*<sup>54</sup> him. He liked no one to be in any way superior to him. He chose his ministers, therefore, not for their knowledge, but for their *ignorance*<sup>55</sup>; not for their capacity, but for their want of it. His *vanity*<sup>56</sup>, his unreasonable desire to be admired, ruined him. His ministers, his generals, his mistresses [...] soon understood this fatal weakness. They praised him and spoiled him, for it was the one way they could approach him. This is why his ministers, drawn from the non-noble class, had so much authority. They had better opportunity to *flatter*<sup>57</sup> him and tell him that all good works came from his actions.

**TASKS B**

First, do the two tasks on your own. Once you've finished task 1 & 2, raise your hand to find a partner to compare and discuss your results.

1. Looking at paragraph 2, **explain** what Duke Saint-Simon criticises about Louis XIV. Write down sentences using words from the box to the right.

Historians often try to explain how things developed = **what led to what** (with the ultimate aim of explaining why the world is as it is today!)

The phrases on the right help you **express causes and effects**. Use them to show that you know how historical circumstances and events are causally linked!

introducing causes	introducing effects/ consequences
because because of (+noun) as since due to (+noun)	so therefore as a result, ... as a consequence, ... consequently, ... ..., resulting in ... ..., leading to .../
X led to Y	

<sup>51</sup> available online:

[http://www.northernhighlands.org/cms/lib5/NJ01000179/Centricity/Domain/58/louis\\_XIV\\_Primary%20source\\_s.pdf](http://www.northernhighlands.org/cms/lib5/NJ01000179/Centricity/Domain/58/louis_XIV_Primary%20source_s.pdf)

<sup>52</sup> Haltung, Manieren

<sup>53</sup> prinzipientreu, mit hohen Grundsätzen

<sup>54</sup> erschöpft, überdrüssig werden

<sup>55</sup> Unwissenheit, Ahnungslosigkeit

<sup>56</sup> Eitelkeit

<sup>57</sup> schmeicheln

**Comment:** It is recommended to specify which type of outcome the learners should produce:

- Asking them to produce full sentences makes sense for cognitively challenging tasks, longer coherent outputs that require good organisation, more sophisticated language and/ or are focused on a CDF type the learners need more practice with.
- Notes or keywords are particularly useful for tasks more prone to oral language/ interaction. Notes are intended to help students participate more readily, especially if they are struggling with fluency.

Another aspect to consider is variation between key words/ oral tasks and full sentences/ written tasks to cater for different learning styles and personality types.

2. Look at the last two lines of paragraph 2. **Give possible reasons why Duke Saint-Simon might have mentioned** that Louis' ministers were "drawn from the non-noble class". Write down your answer in a full sentence using a strategy from the box to the right.

**Comment:** Abstract linguistic input might become clearer to the learners by adding a concrete subject-specific example.

Whenever we talk about something **hypothetical** (= something we don't or can't know) we need to show explicitly that we are **speculating**. We can do this by using modal verbs (**could, would, may, might**). We could also use words expressing probability, such as **probably, maybe, or likely**, or construct **either-or sentences** or **if-sentences**. We can also express uncertainty by saying that something **seems/ appears** instead of saying that something is.

E.g.: Considering what we know about Marie-Antoinette, she really might have said her famous quote "let them eat cake". Yet, there is no evidence, so she probably never said it.



**\*\*Now raise your arm to find a partner to discuss and compare your results! \*\***

### C) Portrait of Louis XIV of France (1702)

*First, do the two tasks on your own. Once you've finished task 1 & 2, raise your hand to find a partner to compare and discuss your results.*

1. **Describe** the picture (colours, light, clothes, accessories, posture, etc.) and identify the type of picture. Take notes (keywords) and look up words you don't know.



**Tip:** When you are asked to *describe*, only describe what you see; do not yet interpret.

2. **Explain** why Louis might be depicted this way, using full sentences. (Remember the language boxes about reasons and hypotheticality)



Figure 1. (Rigault, H. (1702). Louis XIV of France [Oil on canvas]. Retrieved from <https://www.louvre.fr/oeuvre-notices/louis-xiv-1638-1715>)

**\*\* Now raise your hand to find a partner to discuss and compare your results! \*\***

#### **D) Text by a historian<sup>58</sup>**

- 1 Louis XIV (1638–1715) was the longest reigning king in French history. His long rule was a period of dramatic political, social, and cultural development as well as extraordinary turbulence. As a boy, he lived through the last decades of the Thirty Years War and chaotic *civil wars*<sup>59</sup>. After deciding to rule personally, he greatly strengthened the authority of the absolute monarchy, made France the dominant power in Europe, and, as the self-proclaimed Sun King, guided the bloom of classical French culture from his glittering court at Versailles. His last three decades were darkened by great wars, religious *controversy*<sup>60</sup>, *famine*<sup>61</sup>, state bankruptcy, and economic *stagnation*<sup>62</sup>.
- 2 Ever since, historians have struggled with the meaning and significance of his reign. For Voltaire, the age of Louis XIV was an era of cultural achievement. The supporters of the French Revolution *condemned*<sup>63</sup> him as the chief architect of royal tyranny. French historians of the 19th century, strongly influenced by *contemporary currents*<sup>64</sup> of liberalism and nationalism, portrayed him as a great state-builder who laid the *foundations*<sup>65</sup> of the modern state.

**Language tip:** The language of history is usually full of nouns to make texts more **concise** (= *kurz und prägnant*) and **formal**.

In order to have many nouns, historians like to change verb phrases (and sometimes adjectives) into noun phrases. This process is called **nominalisation**.

For example: "The age of Louis XIV was an era in which they culturally *achieved* a lot" can be changed to "the age of Louis XIV was an era of cultural achievement" - as found in this text.

By doing so, actions can be transformed into more abstract concepts and general developments. This way, one does not need to specify who actually did something, like in the example provided here.



#### **TASKS D**

*First, read the language tip. Then do task 1-4 together with a partner.*

1. Have a look at the text and underline at least three of these nominalised phrases.
2. In the second paragraph, different views on Louis XIV are presented. How are these views reported here? **Highlight** these words using different colours.
3. Think of other words you can use to introduce ideas of others and write them down.

---

<sup>58</sup> Dee, D. 2013. Louis XIV, King of France. In *Oxford Bibliographies Online Datasets*.  
<https://doi.org/10.1093/obo/9780195399301-0182>)

<sup>59</sup> Bürgerkriege

<sup>60</sup> Auseinandersetzung

<sup>61</sup> Hungersnot

<sup>62</sup> Stillstand

<sup>63</sup> verurteilen

<sup>64</sup> zeitgenössische Strömung/ Bewegung

<sup>65</sup> Fundament, Grundlage

4. **Report** how the author of this text **evaluates** Louis' reign. First, take notes (keywords) on your own. Then talk to your partner and try to use reporting words to indicate the author's views.

**Comment:** This task is quite language-focused (nominalisation & REPORT), making it somewhat unusual for learners of history. Thus, the procedure is scaffolded in very small steps, guiding the learners through the process.

Furthermore, nominalisation is only briefly introduced, mostly to raise awareness. No productive task centring on nominalisation is added to avoid rejection by the learners. Previous results suggest that (1) nominalisation is not a core issue and (2) too much linguistic focus without clear connection to the content is not welcomed by learners of history. While it is difficult to create tasks that genuinely connect nominalisation and content, it is still a common feature of historical discourse. To subtly build up attention towards this feature nonetheless, awareness-raising remarks and brief exercises are included from time to time.

*\*\* Now raise your hand to find a partner to discuss and compare your results! \*\**

**FINAL TASK:** Now that you've read, heard, and seen a lot about Louis XIV, how do you **assess** his reign and historical significance? Consider the tips in the boxes and write down 3 to 5 sentences.

*\*\*Once you are done, raise your hand again to find a partner. Read each other's texts and then discuss to what extent you agree. \*\**

Tip: Always distinguish between fact, opinion, or interpretation and use language to mark whether something is **fact**, *opinion*, or interpretation.



✓ e.g., **we know from other sources/ historians seem to agree that/ it is a well-known fact that...**

✓ *From my point of view/ I am of the opinion/ I believe, etc.*

✓ Looking at this quote, it seems that/ we can assume/ this indicates that/ this could mean...

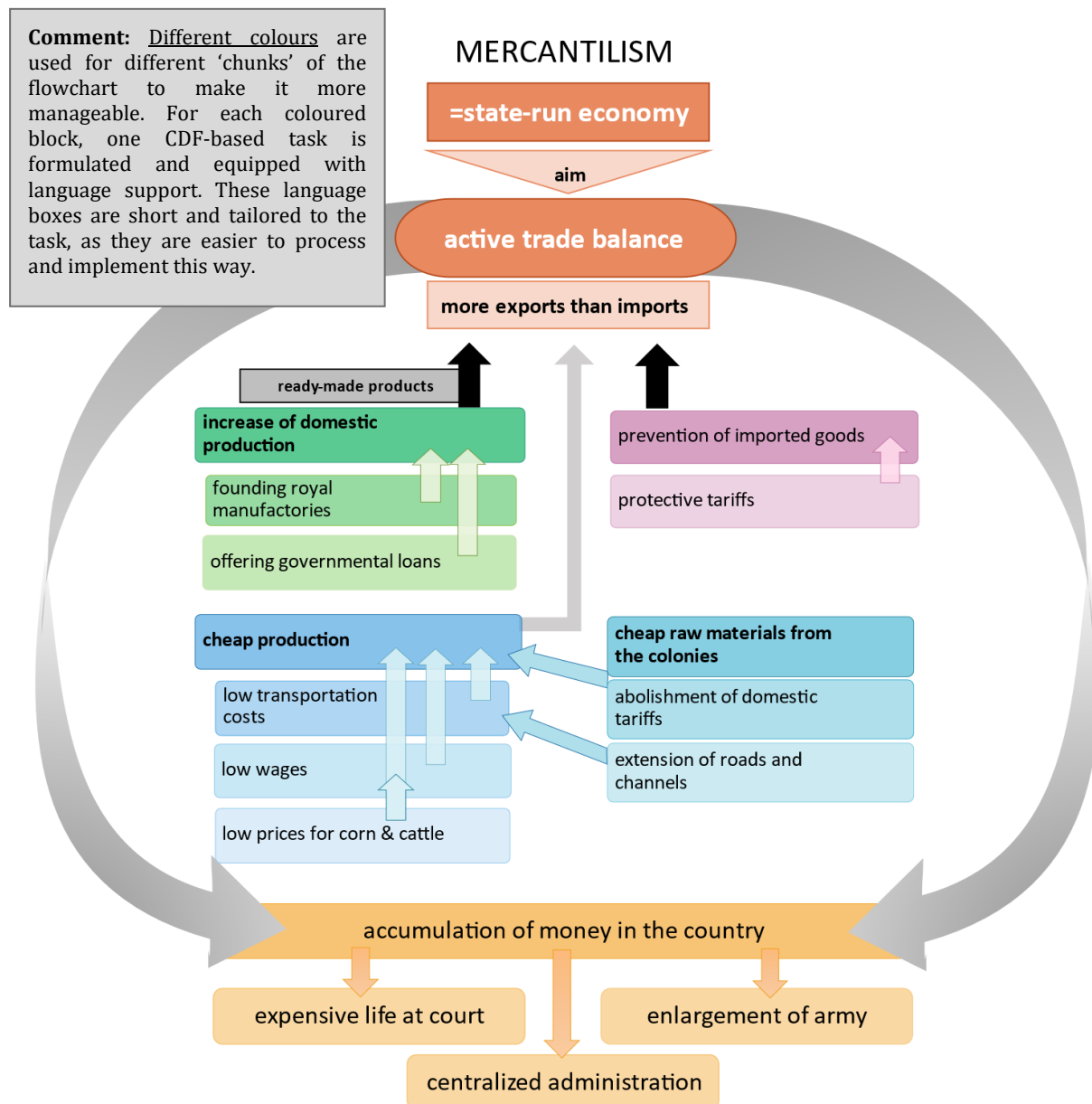
→ Show that your answer is not the only interpretation! You can use modals like could or would, tentative words (e.g., seem, assume, suggest) or you could add adverbs/ adjectives (probably, possible etc).



**Comment:** These tips are based on the learners' difficulties as identified in the pre-intervention task results of the pilot and the main study, which are, in this case, precision of expression and hedging. To facilitate the learners' implementation of these tips, concrete examples and phrases are provided.



Have a look at the flowchart<sup>66</sup> and do the tasks below. There will be language boxes and tips to help you.



- 1) Looking at the orange part at the top, write a **definition** of mercantilism, similar to the examples in the box to the right.

#### Components of a definition:

term = **broader category** + *specifics/ description* (+ *further information/ limitations/ examples*)

e.g. The term democracy refers to a **system of government**, in which power is either held by elected representatives or directly by the people themselves and which is based on the belief in freedom and equality between people.

<sup>66</sup> The flowchart is partly based on and translated from Scheucher, A., Ebenhoch, U., Staudinger, E., & Scheipl, J. (2013). *Zeitbilder 6 [images of history 6]: Geschichte und Sozialkunde/Politische Bildung [history and social studies/ political education]* (1<sup>st</sup> ed.), (pp. 47-48). ÖBV.



- 2) Green: Now **explain** how an increase of production can be achieved. Write at least one sentence and use words that signal which factors are causes and which are effects.

**phrases for cause and effect**

Z resulted from X  
Z was caused by X  
X, resulting in Z.  
Z (passive voice) through X  
X, Y. Therefore, Z.

- 3) Purple: Next, **explain** how protective tariffs contributed to an active trade balance. Formulate at least one sentence.

**phrases for cause and effect**

X led to Z  
Z was a result of X and Y  
One/ Another reason for Z was X  
X, contributing to Z.  
X, Y. As a consequence, Z.

**Tip:** Always make sure that these phrases fit to the content (both in terms of meaning and grammar).

E.g., when you use "because", make sure you really present reasons.

- 4) Blue: Now, **explain** how cheap production was achieved in mercantilism in full sentences. This time, you have more than one factor. Make sure you structure your answer in a comprehensible way.

**phrases for ordering**

First of all,  
Firstly, secondly, ...  
Another factor in...  
Another way of...  
Finally,

**Comment:** Now that the learners have had some practice EXPLAINING, another layer is added, namely multiple factors in need of organisation (REPORT).

**Tip:** First say that there are a couple of factors and then structure them logically.

- 5) Yellow: **Report benefits** of an active trade balance, using phrases from the box below.

**phrases to express positive outcomes**

Z allows X/Y  
Z permits X/Y  
Z makes X and Y possible  
Z ensures X and Y

- 6) Now, have a look at the whole flowchart and our summary again. Is there anything in terms of content that you don't understand?

**Comment:** Again, a brief remark pointing their attention towards nominalisation and its effect on the text are included to subtly raise awareness in this regard.

**Look again:** By using the nouns from the flowchart, you've created a formal and concise text!

- 7) You might have noticed that the flowchart only contains benefits. First, take notes on your own and then **discuss** with a partner (actual and potential) negative consequences of a mercantilist system.

## 7.2.3 Unit II: the Industrial Revolution (cycle 3)


### 7.2.3.1 Design process of unit II

For designing unit II, T<sub>A</sub> and I met two times, once to devise a general outline and once to discuss and improve what had been drafted based on the initial discussion. In our first meeting in early April, we agreed on main themes, the scope, and the sequence of sub-topics in line with the curriculum and the teacher's annual programme. It was decided that the unit should start off with working through textual input on the origin of the Industrial Revolution (IR). Next, population development and urbanisation would be covered by working with a line chart. Subsequently, living conditions during the IR should be discussed with the help of visual sources, while the topic of urban development and housing should be based on illustrative textual sources. Then, working conditions, including child labour, should be covered in a more extensive and creative task that involves a variety of sources. As in the previous unit, the aim was to create a unit centred on historical sources in a way that enables learners to critically engage with them by providing linguistic support and CDF-based scaffolding, reflecting the curricular aim of “us[ing] historical **sources** critically to **reconstruct** and **deconstruct** history” (Austrian Federal Ministry for Education, 2014, p. 91). Topic-wise, this unit focuses on “**describ[ing]** the **influence** of historical developments on individuals, the society, and the state” and “significant historical processes of change” as well as on “**analys[ing]** and **explain[ing]** their causes” (Austrian Federal Ministry for Education, 2014, p. 91). Finally, a whole-group discussion should allow for a comparison to and reflection on current industrialisation processes and conditions of labour, which is in line with the curricular aim of “**present[ing]** and **analyz[ing]** social developments and **assess[ing]** their importance in historical context” (Austrian Federal Ministry for Education, 2014, p. 92).

Our starting points for the unit were materials we already knew from our previous teaching and research experiences (see Bauer-Marschallinger, 2016). With our aim in mind, we went through these materials, also considering previous research cycles and the target group. In the process, we selected a number of historical sources for the individual sub-topics, discussed and developed potential corresponding tasks or possible ways of adapting existing tasks for the purpose of this study, similar to the materials of unit I. Based on what was decided in this session, the lesson plan and the tasks were formulated, which were later reviewed by an experienced researcher. Then, T<sub>A</sub> and I met mid-April to improve and finalize our design and to organize the implementation.

One example of an innovation of the third cycle that stem from previously gained results are the newly introduced *impulse questions*. The impulse questions intend to initiate deeper reflection and, like in Figure 18, establish connections to the present and the learners' world of experience, which is something the learners asked for in the interview. The impulse question presented below further considers an issue observed in connection to item 2 of the pre- and post-tasks, which required the learners to explore the motives of those that created the historical source in question. Here many answers were variations of “to show us how it was”. As already argued in section 4.1.5,

such answers are characterized by presentism (Carretero & van Alphen, 2014), lacking awareness of the historical context of production as well as critical reflection.



**Impulse question:**

When a journalist today takes a picture, do you think (s)he only wants to show future generations what it was like in 2019? Together with a partner, collect ideas for potential motives for taking pictures other than just depicting what something looks like.

c) Now, **discuss potential reasons** why the artist drew this particular picture the way (s)he did. Answer in full sentences and mention at least two different ideas.

Remember to use **hypothetical** language:

- ☐ *could/ would/ might*
- ☐ *probably/ maybe/ possibly*
- ☐ *another potential motive/ reason...*
- ☐ *if-sentences*
- ☐ *I could imagine that ...*

Figure 18. IR – part 3: impulse question and task c

The task presented above is aimed at foregrounding the context of production by shifting the thought process to their own timeline and inspire their creativity when it comes to motives. In the next exercise (task c), the students should then take these insights as a starting point for an analysis of motives relating to a historical source. To support their production, the learners are provided with input on hypothetical language.

Previous (preliminary) results and impressions from the implementation phase of unit I also informed our decisions on which CDFs should be featured and how much or which language support the individual tasks require. For example, according to the teacher's experiences, the learners still struggled with comparing, which also concurs with the students' self-reports and the results of the pre- and post-intervention tasks:

76	English translation	Original quote
TA Session 4	<p>TA. Well, comparing [...] this is something they have problems with [...] you know, it's always one paragraph on the first source and one paragraph on the second,</p> <p>R: =It was the same in the [post-task]</p> <p>TA: =but it's not connected. [...] this does not only concern this class but all.</p>	<p>TA: <i>Also vergleichen [...] das ist etwas, mit dem sie Probleme haben [...] Weißt, immer so ein Gsatzl erste Quelle, ein Gsatzl zweite Quelle, aber nicht</i></p> <p>R: <i>=Das war ja jetzt auch so. Bei dem [post-task]</i></p> <p>TA: <i>=dass das ineinander geht [...] das eben, betrifft nicht nur diese Klasse, sondern alle.</i></p>

The teacher reported that most learners tended to describe two items independently without connecting or contrasting them. Therefore, part 4, task b (see below) focuses on comparing 'then' and 'now' and provides phrases that facilitate comparison rather than independent description of two periods. Furthermore, we agreed to allocate quite a bit of time to this exercise. In the first phase, learners should take notes individually and then the teacher would moderate a whole-class discussion, using the whiteboard to systematize input. During this discussion, the teacher would model comparative structures and ensure that learners use appropriate language.

b) Labour and industrialisation then and now: **Compare** labour and industrial production in the time of the Industrial Revolution to today's different parts of the world. Take notes, considering living and working conditions, child labour, pollution, and population development.

Here are some phrases that could help you for the discussion:

differences	similarities
... while ...	both/ all
in contrast, ...	neither
differ in	similar
in comparison to	share
... whereas ...	... as <i>adjective</i> as ...
compared to	
more/ less than	

Figure 19. IR – part 4b: CATEGORIZE activity

Concerning the issue of diverging paces, T<sub>A</sub> rejected T<sub>B2</sub>'s suggestion of letting learners work in their own time before asking them to find a partner due to social issues in this class. She reported that group A was affected by strong cliques, resulting in their refusal to work with people other than their friends, which was an issue the intervention could hardly counteract. Instead, optional *fast-track* activities were introduced to ensure that different levels of ability and pace were catered for. Since the scaffolding inherent to our pedagogical design is intended to predominately support low to mid achievers, these fast-track activities are designed to be challenging for high achievers. As such, they are more explorative, nuanced, and less scaffolded, focusing on subject-specific linguistic aspects potentially new to them. Figure 20 provides an example:

c) Fast track task: These quotes use rather loaded words (=negativ besetzte Wörter). First, circle these loaded words. Then, **provide reasons** why these people might have formulated it like that.

Figure 20. IR – part 3c: fast-track activity

Here, the learners' attention was pointed towards loaded words, which they might usually not consider when working with historical texts. At the same time, this linguistic nuance functions as a starting point for further reflections.

Apart from devising and refining these changes, we also discussed more practical aspects, e.g., word limits, suitable social formats, the content of the glossary, selection of gaps for the gap-fill task, and the timeline. As for social formats, it was decided to turn more challenging tasks into pair work so that stronger learners could help their peers (e.g., part 3/ textual source/ task b). Relating to time planning, it was considered that parts repeating aspects previously dealt with would not take that much time. Still, cuts were needed to keep to the schedule without overwhelming learners (again). Thus, it was decided to transform the summary task (part 3/ textual sources/ task a) into an oral task, also considering that the following task requires the students to write too. Another strategy was to provide enough vocabulary in the glossary for the textual sources characterised by challenging and archaic lexis. The fast-track activities, too, were intended to allow for more realistic time planning. Finally, we agreed to debrief after each session, deciding whether adaptations to our schedule were needed. At the end of this final design session, the organisation of the implementation phase and the post-intervention written task was dealt with, and it was decided, once again, that I would work in the adaptations based on our discussion, which the teacher would subsequently review.

### 7.2.3.2 Final lesson plan of unit II

grade	number of lessons	language
11/ 3 HAK (secondary colleges of business administration)	4-5	English (German)

#### Themes:

- curriculum 5<sup>th</sup> semester: “Milestones in the historical development: Industrial Revolution” (Austrian Federal Ministry for Education, 2014, p. 92)
- curriculum 6<sup>th</sup> semester: “Changes in the world of work and in social structures through industrialization” (Austrian Federal Ministry for Education, 2014, p. 92)

#### Topic: Industrial Revolution

- origin in England, population development & urbanisation, working and housing conditions (social consequences)

#### Objective:

Curriculum, 5<sup>th</sup> and 6<sup>th</sup> semester (Austrian Federal Ministry for Education, 2014, pp. 91-92):

- **describe** the **influence** of historical developments on individuals, the society, and the state (DESCRIBE, EXPLAIN)
- **describe** significant historical processes of change, **analyze** and **explain** their causes (DESCRIBE, EXPLAIN, EXPLORE, REPORT, EVALUATE)
- **assign** accomplishments of civilization to epochs (CATEGORIZE)
- **present** and **analyze** social developments and **assess** their importance in historical context (DESCRIBE, EXPLAIN, CLASSIFY, EXPLORE, REPORT, EVALUATE)
- use historical sources critically to **reconstruct** and **deconstruct** history (DESCRIBE, EXPLAIN, CATEGORIZE, EXPLORE, REPORT, EVALUATE)

Further, discipline-specific language focus (see also subsection 3.2.2.2):

The students can

- use **nominalisations** appropriately to increase concision, formality, and level of abstractness of their output
- **hedge** their ideas to make their claims less absolute and qualified
- **link** their ideas effectively (appropriate choices and form)

rough time frame	procedure/ content	interaction format	skills / language system	language	historical competence	materials	learning phase	
10'	Task paragraph 1: definition and scope of the IR	individual	reading writing	DF, EV	FC	Text + WS part 1	introduction	
15'	Task paragraph 2: development of the IR	individual notes, then pairs	reading speaking	DS, RE, EA, describing change	MC + FC		elaboration	
15'	Task paragraph 3: debate on the time frame of the IR (+ optional fast-track task)	individual, then comparison of task 1-3 in plenum	reading writing	RE, EV, reporting verbs	MC + FC		elaboration (advanced elaboration)	
10'	Task paragraph 4: Why was Britain first?	pair, then comparison in plenum	reading speaking	nominalisation	FC		elaboration	
10'	Continuation task paragraph 4: Why was Britain first?							
30'	1) Population development and urbanisation Task a: describing population development	individual, then comparison in plenum form	writing speaking	DS, graph description	MC	WS population development (IR part 2)	elaboration	
	Task b: comparing population development of different groups			DS, CA, graph description				
	Task c: defining urbanisation			DF	FC			
	Optional fast-track task d: potential consequences of urbanisation and population growth	individual	writing	EO, EA	FC		advanced elaboration	
10'	2) Urbanisation and living conditions: Task a: describing a picture	individual	speaking	DS	MC	WS living conditions (part 3)	elaboration	
20'	Task b: analysis of visual source		writing	DS, EV	MC		elaboration	
	Impulse question		pairs, then comparison in plenum	speaking	EA			OC
	Task c: potential motives of the artist				EA, EO, hypothetical language, hedging			MC
30'	Living conditions: textual sources Task a: summarizing multiple sources	individual, then plenum	reading speaking	RE, DS	MC		elaboration	

	Task b: reasons for bad living conditions	pairs	speaking	EA	MC		elaboration
15'	Task c: optional for fast students: loaded words	individual	reading	EO, EA, loaded vocabulary			advanced elaboration
	Impulse question	pairs	speaking	EA, EO	OC		reflection
35'	3) Working conditions: writing a report	individual	reading writing	RE, EV, DS	MC	WS working conditions (part 4)	elaboration
15'	Continuation of report	individual	reading writing	RE, EV, DS	MC		elaboration
35'	Comparing labour and industrial production then and now (bring in industrialisation 4.0 → artificial intelligence) One could split the board into 'then' and 'now' and keep track of the students' points that way	individual, then plenum	speaking	CA, EV, EA, DS, RE	OC	WS part 4, whiteboard	transfer and reflection

### 7.2.3.3 Final version of worksheets of unit II

In the following, the four parts of this unit are presented. Part 1 deals with the origins of the IR, part 2 covers urbanisation and population development, part 3 discusses living conditions resulting from the changes described in part 2, and part 4 deals with the working conditions as a result of industrialisation in the past and now.

Like for unit I, space provided for students' answers has been reduced and comments outlining some of the decisions that informed the design have been added in grey boxes. Comments that were already included in 7.2.2.4 are not repeated. Further materials, including drafts and revisions as well as the teacher's version containing a key and practical tips for the implementation and adaption, can be found in the appendix repository, [section III/B \(unit II – Industrial Revolution\)](#).

## Text: The Industrial Revolution

Part I: Read through the text and do the tasks on the worksheet.<sup>67</sup>

### 1 What was the Industrial Revolution?

**Comment:** Sub-headlines help learners understand the overall topic of a section.

The Industrial Revolution was a period from the 18th to the 19th century where major changes in agriculture, manufacturing, and transport had a far-reaching effect on the socio-economic and cultural conditions starting in the United Kingdom, and then subsequently spreading throughout Europe, North America, and eventually the world. The onset<sup>68</sup> of the Industrial Revolution marked a major turning point in human history. Almost every aspect of daily life was eventually influenced in some way, from where people lived and worked to how people viewed the world and their life expectations.

### 2 How did it develop?

Starting point: 18<sup>th</sup> century, Great Britain:  
*manual labour*<sup>69</sup> & *draft-animal*<sup>70</sup> agriculture → machine-based manufacturing and agriculture

Important developments:

- mechanisation of textile industries
- development of iron-making techniques
- increased use of refined<sup>71</sup> coal
- new canals, improved roads & railways → trade expansion
- introduction of steam power → increase of production

19<sup>th</sup> century: spread throughout Western Europe and North America

ever since then: industrialisation all over the world

**Comment:** The phrases underlined are phrases that are useful for the tasks (current and upcoming) or which were introduced in the previous unit.

By the way:  
Useful  
phrases are  
underlined  
in grey.

**Comment:** Note that the bullet points – and the text in general – contain many nominal phrases. This should model appropriate language and support target-like production.

### 3 When exactly did it happen?

The period of time covered by the Industrial Revolution varies with different historians. Some assert that it 'broke out' in Britain in the 1780s and was not fully felt until the 1830s or 1840s, while others believe that it occurred roughly between 1760 and 1830. In contrast, other historians argue that the process of economic and social change took place gradually and the term revolution is not a true description of what took place. This is still a subject of debate among historians.

**Comment:** As guessing vocabulary from context is a useful skill (and one that is tested in the final English exam), we did not include many words in the glossary in section 4. Furthermore, a receptive task like this also allows for a less comprehensive glossary.

<sup>67</sup> This is a shortened and adapted version of a text provided by *International School History*. (2014). "Unit 5: What were the social consequences of the Industrial Revolution?" Retrieved from [http://www.internationalschoolhistory.net/eeb3/s5/extra/social\\_consequences\\_ppt.htm](http://www.internationalschoolhistory.net/eeb3/s5/extra/social_consequences_ppt.htm)

<sup>68</sup> Beginn

<sup>69</sup> Handarbeit

<sup>70</sup> Zugtiere

<sup>71</sup> verfeinert



#### 4 Why was Britain first?

Unlike with the time frame, economic historians agree that the Industrial Revolution began in Great Britain due to a number of reasons:

- **Geography & climate:** Britain is a small country with many navigable rivers, good mineral deposits and relatively few huge natural obstacles to \_\_\_\_\_. Furthermore, there is plenty of water that can be used for waterpower, which, in turn, is crucial for \_\_\_\_\_. As for climate, Britain offers a diverse, yet mild climate, ensuring fruitful and stable agriculture.
- **Earlier economic development:** The 18<sup>th</sup> century had seen remarkable advances in both trade and industry, \_\_\_\_\_. Moreover, Britain's active foreign policy gave rise to frequent boosts to the iron and textile industry.
- **Rising population growth:** Britain's population almost doubled in the course of the 18<sup>th</sup> century. More people mean more opportunities for \_\_\_\_\_.
- **The advantage of empire:** Britain had a healthy lead over European competitors in overseas trade, particularly in India and the Americas. In addition, France, which was England's greatest opponent within Europe, was economically damaged by war and \_\_\_\_\_.
- **The role of government:** In comparison to most other European countries, Britain was a \_\_\_\_\_ and, at that time, relatively politically stable yet open to change and innovation.
- **British society:** Some features of the British society of the 18<sup>th</sup> century also contributed to Britain's head start in industrialisation.

First of all, British society at the time was known for its \_\_\_\_\_. The famous industrial innovations – the steam engine, *Power Loom*<sup>72</sup> and other inventions – were symptomatic of a much wider commitment to experiment and innovate. The development of transport networks helped to spread new scientific and innovative ideas. What is more, knowledge could be shared as a result of the \_\_\_\_\_.

Secondly, Britain is a Protestant country. Max Weber, a famous *sociologist*<sup>73</sup>, argues that Protestant values and lifestyles were more conducive to<sup>74</sup> \_\_\_\_\_ than Catholic values.

Thirdly, Britain was considered an 'open society'. According to a number of historians, the 18<sup>th</sup>-century British society was comparatively liberal, and thus talented people could rise to \_\_\_\_\_.

---

<sup>72</sup> Maschinenwebstuhl

<sup>73</sup> Soziologe: jemand der Gesellschaft beforscht

<sup>74</sup> förderlich für

## Tasks: The Industrial Revolution

### (Industrialisation part 1)

#### Paragraph 1:

**Explain** the view that “the Industrial Revolution was much more than a revolution in industry”.

First, take some notes, then discuss with a partner.

EXPLAIN means that you have to provide reasons and make it obvious that you are presenting reasons here by using appropriate phrases. Remember: There's more than *because*...

**Comment:** Since learners seem to struggle with justifying their views, this box was included although a similar box was already included in unit I. Furthermore, the note on “because” was added, as the students appear to overuse it, also in inappropriate contexts.

**Paragraph 2:** In your own words, **outline** in full sentences the development of the Industrial Revolution. Use some phrases from the box below:

phrases indicating change	describing consequences
to (gradually) replace	as a result, ...
give rise to/ lead to	as a consequence, ... / consequently, ...
to radically change/ transform; a radical change/ transformation	therefore, .../ thus, ...
to increase; increase of	..., resulting in ...

**Comment:** These phrases were included since the data has shown that learners struggle with indicating change and expressing cause and effect.

**Paragraph 3:** **Report different views** on the exact time frame of the Industrial Revolution and **explain why** some historians have been unhappy to use the term “revolution” for the industrial changes in the late 18<sup>th</sup> and early 19<sup>th</sup> century.

Use reporting phrases whenever you refer to other people's views:

'strong' verbs	neutral phrases	'tentative' verbs
argue	state explain	suggest
assert	report point out	assume
claim	according to	propose

**Comment:** This orange box here should highlight how the language box relates to the subject of history and also includes explicit instruction on what to look out for (including an example).

In history, when we report other people's views, we:  
1) need to show that this **is** somebody else's view (by using reporting phrases)  
2) should indicate **how strongly** this person argued their view.

**Attention:** Reporting verbs also indicate **to what extent YOU agree** with these views. E.g., *claim* suggests that you are sceptical about the reported views.

### Fast-track task:

**Comment:** Fast-track activities are for those students who finish earlier than their peers. Since the scaffolding inherent to the pedagogical design of this unit is intended to predominately support low to mid achievers, these fast-track activities are designed to be challenging also for high achievers. As such, they are less scaffolded, more explorative and nuanced, focusing on linguistic aspects of historical discourse that might be new to them.

Circle in all reporting verbs in this paragraph and assess how strong they are and to what extent the author of this text seems to agree with the reported view.

Paragraph 4: Read through the different factors of Britain's advantage in the industrialisation process and **fill in the gaps** with the words from the box below. Look up words you don't know. You can work together with a partner.

A. constitutional monarchy	B. an unsustainable, mercantilist economy	C. production and consumption	D. hard work and accumulation of money
E. movement of trade and people	F. inventiveness	G. technological development	H. wealth, influence, and power
I. increased availability of newspapers and magazines	J. resulting in rising incomes and spending power		

**Comment:** The gaps are selected purposefully: Either they represent central concepts (from a content perspective) or nominalised phrases, guiding their attention subtly towards this feature of historical discourse. When selecting the gaps, we also tried to make sure not to include too many gaps, as there needs to be enough context for the learners to understand the connections.

## Population development during the time of the Industrial Revolution (Industrialisation part 2)

1) **Population development:** With the help of the graph<sup>75</sup> and the boxes,

- describe** the total population development of England and Wales between 1751 and 1901.
- compare** the development of the rural and the urban population of England and Wales between 1751 and 1901.

**Comment:** When learners are asked to compare, they tend to describe the two (or more) items independently without properly connecting them. Using some of the phrases provided in the right column of the box might 'force' the learners to link the two.

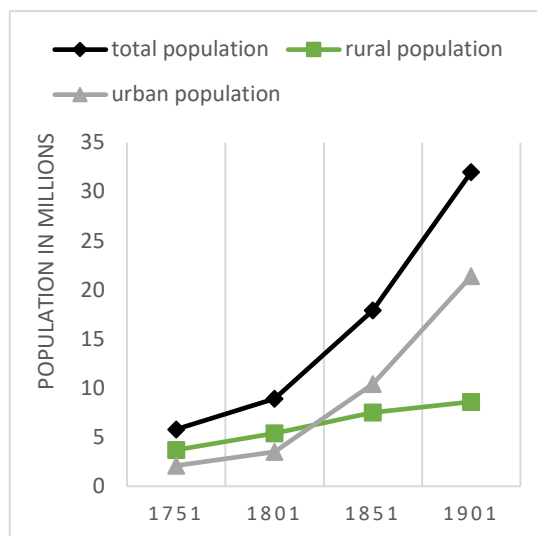
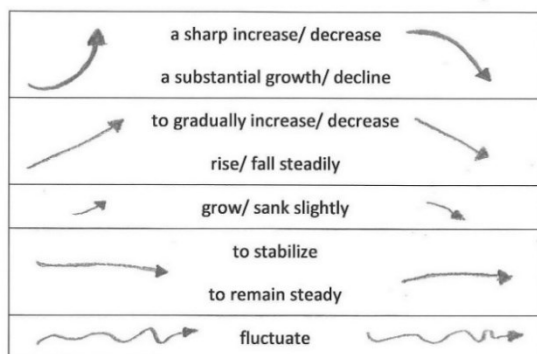


Figure 1. Population development in England and Wales, 1751-1901

General phrases	Comparing trends
According to this graph, ...	while
This graph shows/ illustrates...	whereas
From 1751 to 1801 ...	compared to
Between 1801 and 1901 ...	in contrast
During the period from ... to	unlike

- Define the term urbanisation.

Remember that **a definition consists of 3 parts:**

the broader category + defining feature + further specification/ example

**Comment:** This language tip was already introduced in unit I. Still, a (shorter) repetition could support consolidation.

- Optional fast-track task:** Discuss potential consequences of urbanisation and population growth for housing and living conditions.

**Comment:** This is a lead-over to the next part, inviting learners to explore and reflect on the topic before being guided through it, ensuring that faster learners can engage with higher-order thinking skills.

<sup>75</sup> The graph was created based on data provided here: <https://1841census.co.uk/1570-1750-estimated-population/> and [http://www.visionofbritain.org.uk/census/GB1841ABS\\_1/6](http://www.visionofbritain.org.uk/census/GB1841ABS_1/6).

## Urbanisation and living conditions (Industrialisation part 3)

VISUAL  
SOURCE:



Figure II. A poor household in Bethnal Green, London, in 1836. Retrieved from <https://historyatnormandale.wordpress.com/2017/04/07/living-conditions-and-urbanization/>


- a) **Describe** the picture (people, furnishing, colours). Here, you are welcome to only use keywords.

**Comment:** Content-related keywords were added to specify the target outcome, making the task less vague. In the interviews, such specifications were appreciated by the students and teachers, as they keep the lessons focused.

ARGUE means that you provide reasons for your view. Again, use language to clearly show which part refers to your view and which part are reasons for said view.

For example: *It seems to me that X since/because, etc.*

- b) **Argue** what kind of mood this drawing creates. Again, keywords are enough.

 **Impulse question:** When a journalist today takes a picture, do you think (s)he only wants to show future generations what it was like in 2019? Together with a partner, collect ideas for potential motives for taking pictures other than just depicting what something looks like.

**Comment:** The impulse questions intend to initiate deeper reflection and establish connections to the present and the learners' world of experience.

Perceiving a historical source in its historical context and understanding underlying motives is something many learners struggle with. By transferring the thought process onto their own timeline, they might be more ready to apply this kind of investigation when working with historical sources.

- c) Now, **discuss potential reasons** why the artist drew this particular picture the way (s)he did. Answer in full sentences and mention at least two different ideas.

**Comment:** The learners now have to apply a similar analysis to the historical source, supported by some phrases.

Since using hypothetical language is quite challenging for many learners, this box provides not only an awareness-raising remark but also concrete phrases once again (despite their inclusion already in unit I).

Remember to use **hypothetical** language:

- *could/ would/ might*
- *probably/ maybe/ possibly*
- *another potential motive/ reason...*
- *if-sentences*
- *I could imagine that ...*

## TEXTUAL SOURCES:<sup>76</sup>

**Comment:** Although the vocabulary is quite challenging, the quotes are still appropriate due to their highly illustrative character. The glossary and the scaffolded tasks are intended to support their understanding.

- 
1. Medical doctor in Manchester in 1820:

*"Whole streets, unpaved<sup>77</sup> and without drains or main sewers<sup>78</sup>, are worn into deep ruts<sup>79</sup> and holes in which water constantly stagnates, and are so covered with refuse<sup>80</sup> and excrement<sup>81</sup> as to be impassable<sup>82</sup> from depth of mud and intolerable stench<sup>83</sup>."*

---

2. Contemporary witness in Bradford in 1840:

*"These towns have been built by small speculators with no interest for anything except immediate profit. A carpenter and a brick-layer club together to buy a patch of ground and cover it with what they call houses."*

---

3. George Weerth, in 1846:

*"In Manchester the air lies like lead<sup>84</sup> upon you; in Birmingham it is just as if you were sitting with your nose in a stove pipe<sup>85</sup>; [...] In Bradford, however, you think you have been lodged<sup>86</sup> with the devil incarnate<sup>87</sup>."*

---

---

<sup>76</sup> <https://schoolshistory.org.uk/topics/medicine-through-time/public-health-in-the-industrial-revolution/bradford-health-in-the-1840s/>  
<http://www.historyhome.co.uk/peel/p-health/recreat.htm>

<sup>77</sup> ungepflastert

<sup>78</sup> Kanalisation

<sup>79</sup> Furchen

<sup>80</sup> Abfall

<sup>81</sup> Exkremente, Kot

<sup>82</sup> unpassierbar, unbegehrbar

<sup>83</sup> beißender Gestank

<sup>84</sup> Blei

<sup>85</sup> Ofenrohr

<sup>86</sup> untergebracht, wohnen

<sup>87</sup> dem Teufel selbst

- a) First, **highlight keywords** that tell you something about 19<sup>th</sup>-century living conditions in Britain. Consider sanitation, hygiene, construction of houses/streets, and pollution.

**Comment:** This first step should help them deal with these sources despite the unfamiliar, historical language and give them a starting point for their summary. One could also suggest using different colours for different categories in order to see what is most central, guiding them even more through the process.

- b) Then, take some notes to prepare for an oral **summary** of 19<sup>th</sup>-century living conditions, based on these sources.

summarizing & highlighting	organizing ideas
generally speaking	first of all
in essence	additionally
essentially	another central feature
in brief	closely connected to (this)

SUMMARIZE means that you briefly recap the key aspects in a systematic fashion → start with the most important and remember to use phrases ordering these aspects.

- c) Together with a partner, **explain why** British cities looked like this in the 19<sup>th</sup> century. (Again, you are welcome to only use keywords.)

**Comment:** Since the texts they work with are quite difficult and the tasks requires the learners to connect topics dealt with in other lessons, which might be difficult for some, pair work is recommended here so that stronger learners can support their peers.

- d) Fast-track task: These quotes use rather loaded words (= negativ besetzte Wörter). First, circle these loaded words. Then, **provide reasons** why these people might have formulated it like that.

**Comment:** This optional task deals with linguistic choices that these students have not really focused on so far. It is intended to encourage (gifted) learners to also consider concrete expressions and linguistic evidence in their source analysis.



#### Impulse question:

If living conditions were this bad, why do you think people still moved to cities? **Provide reasons** for your views.

**Comment:** This impulse question serves as a bridge to the next part: labour. Here, no language box is provided as this speech function has been addressed already in this unit (and the previous). Repeating linguistic input too often might lead to rejection by the learners. In this study, some of the more proficient students perceived frequent repetitions as patronizing.

## Working conditions (Industrialisation part 4)

- a) Imagine you are a factory inspector in the 1830s in England and you are asked to come up with a report on working conditions in British factories. You've interviewed a number of people and gathered the information below. Now, you should **summarize your findings in a report** to the government, in which you

OUTLINE  
works  
similarly to  
SUMMARIZE

- **outline** current working conditions in terms of working hours, health hazards<sup>88</sup> and child labour
- **argue** which changes should be made concerning working conditions

Things to consider:

- Remember, you are working for the government. Therefore, you should use **formal, tentative<sup>89</sup>, and polite language**, especially when you suggest something. You could use phrases like: *to my mind, it seems to me, based on my research I would suggest*, etc.
- Try to use phrases for **ordering and structuring** your report.
- Always be clear **which source you are referring to**.
- Distinguish between **fact, opinion, or interpretation**.
- Use **language to mark** whether something is fact, opinion, or interpretation.

**Comment:** These are all issues that learners seem to struggle with (according to the data collected). At this point, these aspects are repetitions or slightly altered versions of previously introduced language tips (unit I and/ or II). As such, this task forms a consolidation of what they have done throughout the units – also considering that they are now supposed to work with a number of sources and tie in their analyses into a report, revising some of the operations that they had to do during the units.

Write about 100-130 words on an extra sheet or in a new document. Look up words you don't know.

**Comment:** To allow for more learner differentiation, the teachers could let the learners decide whether they want to use the sources provided or do some research on their own to find (additional) sources.

SOURCES:

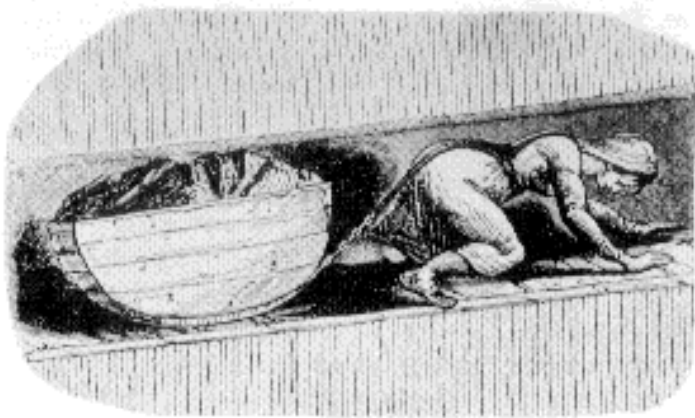
A. *"I began work at the mill I Bradford when I was nine years old ... we began at six in the morning and worked until nine at night six days a week."* Hannah Brown, interviewed in 1832

B. When asked what overseers would do when workers slowed down, Elizabeth Bently, aged six, said: *"Strap (=verprügeln) us ... [The girls] have had black marks on their skin many a time, and their parents dare not to come to him about it, they were afraid of losing their work".*

<sup>88</sup> Gesundheitsgefahren

<sup>89</sup> = Do not present your views as absolute truths but more softly, e.g., by using modals or the phrases mentioned.





C. Girl pulling a coal tub in mine. Picture included in an official report by the parliamentary commission, retrieved from <http://www.victorianweb.org/history/ashley.html>

E. Leonard Horner, a factory inspector, around 1840: *"She was caught by her apron (=Schürze), which wrapped around the shaft. She was whirled round and repeatedly forced between the shaft and the carding engine. Her right leg was found some distance away."*



D. Photograph of Widnes in the late 19<sup>th</sup> century, retrieved from <https://commons.wikimedia.org/w/index.php?curid=2031783>

F. *"The easiness of the work makes long hours possible. Most of the work is merely that of watching machinery and piecing the threads that break."* Nassau Senior, factory owner in 1833

Textual sources were retrieved from (7 April 2019):

- <https://www.historylearningsite.co.uk/britain-1700-to-1900/industrial-revolution/children-in-the-industrial-revolution/> (Source A)
- <http://www.collaborativelearning.org/cottonmill.pdf> (Source B & E)
- <https://www.propofs.com/quiz-school/story.php?title=source-h-the-factory-owner> (Source F)

b) Labour and industrialisation then and now: **Compare** labour and industrial production in the time of the Industrial Revolution to today's different parts of the world. Take notes, considering living and working conditions, child labour, pollution, and population development.

Here are some phrases that could help you for the discussion:

**Comment:** Often, extrovert and/ or strong students blurt out the most obvious aspects before shy or weaker students can even decide whether they raise their hand or not. Therefore, it makes sense to give some preparation time to everybody so that shy students (and students who rather think something through before saying it in class) also have a chance to contribute in the whole-class discussion.

differences	similarities
... while ...	both/ all
in contrast ...	neither
differ in	similar
in comparison to	share
... whereas ...	... as <i>adjective</i> as ...
compared to	
more/ less than	

## 7.3 Implementation phase

This subchapter outlines the implementation of the interventions. Given the amount of data that shed light on both the processes and products of the designs involved in this project, this subchapter is not concerned with a focused analysis of the videotaped and transcribed lessons.<sup>90</sup> Instead, it offers descriptions of the implementation phase of each cycle to increase auditability and transparency as well as to provide a complete representation of the development process. Moreover, these accounts present elements of the implementation that appeared central to the researcher, and as such this subchapter intends to make the researcher's perceptions intersubjectively comprehensible. Finally, this subchapter provides context for the episodes discussed in subchapter 7.4, where the perceptions of the participants involved are corroborated with relevant episodes from the transcripts. To better follow the reports below, the reader may consult the lesson plans and materials available in subchapter 7.2 or [appendix section III](#).

### 7.3.1 Cycle 1

The materials developed in cycle 1 were implemented in school A over the course of five lessons in February 2019 (14<sup>th</sup>, 15<sup>th</sup>, 21<sup>st</sup>, 22<sup>nd</sup>, 28<sup>th</sup>).

In the first lesson, 17 students were present. The lesson started with a whole-class discussion about absolutism with the help of prepared slides to connect to previous knowledge and provide all learners with sufficient working knowledge for the upcoming tasks. The students' contributions were usually short, mostly limited to individual phrases or short clauses. A few students (ELH01, ARJ01, ORH09, LES02, SAA03), however, were highly involved and also produced more complex and longer utterances. The teacher usually commented the learners' contributions in regard to content and, in a few cases, concerning pronunciation, lexis, or grammar in the form of explicit corrections and recasts. One student (HIP11) asked about grammar ("could I say to show how much he is worth? Is that grammatically correct?") which then, interestingly, resulted in a discussion of content (TA: "grammatically yes, but I wouldn't say so [...] simply to show his power, that's what we would say actually in this context"). In the course of the lesson, the teacher tried to involve more students, and consequently the lesson became more interactive. Some students took notes and others only listened. In general, there were only few instances of German and most of these switches related to terminology. Overall, the teacher spent considerably more time on this step than initially planned. In the interview, the teacher later argued that she had taken more time to establish working knowledge because the learners had shown considerable gaps of knowledge. Another reason might be that this mode of teaching, i.e., whole-class teaching, reflects her teaching style more closely than working with source-based tasks as could be observed prior to the project. After this rather long introduction to the topic, the teacher distributed the worksheet on Louis XIV of France at the end of the lesson. She briefly

---

<sup>90</sup> An in-depth and comprehensive analysis of these lesson transcripts is planned for future publication.

argued why working with sources is important and foreshadowed the content of the upcoming lessons.

The second lesson took place the next day when 18 students were present. First, the teacher explained the tasks without setting a time frame, and then the students started to read silently. After a while, some started whispering and talking while working on the sources. Here, it became apparent that prompts like “take notes” or “discuss with a partner” were not always clear (e.g., who takes the notes, what is the sequence?). A couple of students did not seem to discuss at all which is why the teacher reminded them that they were supposed to talk to their partner. When working on the tasks together, some learners mostly used English, while others switched languages more often. Learners prompted each other words and linking devices, and they also co-constructed sentences. Process management was usually done in German, whereas content-related discussions tended to be in English. During this phase, the teacher walked around and provided feedback on both content and language (e.g., “it’s never good to start with ‘to sum up’”). One pair asked the teacher a language-related question, and another pair asked content-related questions out of curiosity. This part of the lesson also indicated great differences between individual working paces (e.g., EVA02 was done quickly, while UYA06 never finished). In between, the teacher also asked whether they still needed time, which some of them confirmed. After 30 minutes, the teacher stopped this phase to compare the students’ answers in class. The teacher elicited answers and commented on content and language. She especially paid attention to linking devices, asking learners to “please use another phrase” or “to use one of the expressions from the box please”. When students went off on tangents, the teacher cut them off and asked other students. Most of the tasks could be solved by the students, sometimes with the help of other students or the teacher’s verbal scaffolding. Especially providing reasons for their claims and expressing evaluations turned out to be difficult for many. Here, the teacher had to prompt them more often and guide through the process more explicitly. Only few students, like ORH09 and ARJ01, managed on their own. The only task that seemed cognitively challenging for everybody related to hypothetical arguments.

In the third lesson, 18 students were present. It started with a quick revision of the previous lesson and then the teacher gave them ten minutes to work on the second source (B) on their own. In between, the teachers asked them whether they found this text difficult to understand, which they negated. While students were working on the source, the teacher was walking around, providing feedback, reminding them of the language boxes and answering questions (for example, EVA02 asked what “drawn from the non-noble class” meant, or IJT12 asked about linking devices). The learners appeared focused and completed the tasks rather silently (especially in the beginning). Some finished really fast, whereas others needed the whole ten minutes. When comparing the results in class, the teacher also asked those that did not raise their hand. Some of them used too many linking devices, overcomplicating sentences. In these cases, the teacher commented on that and guided the learners towards an appropriate reformulation. One aspect that was not clear for

many was task B/b, namely whether they had to find out reasons why Louis chose non-noble ministers or why Duke Saint-Simon mentioned that. In any case, the learners did not use many hedging devices although this task was accompanied by a language box dealing with this aspect. Then, the learners were given another ten minutes to prepare source C. This time they talked more from the start while taking down notes too. Again, the teacher walked around, providing help and feedback, checking what they were writing down and answering the students' questions (about lexis or tenses, for instance). Some pairs did not work in a focused way, like ORH09 and ATN11. Others preferred to work alone on this task. During the comparison phase, most learners wanted to contribute. Here, like usual, the whole-class discussion followed a typical IRF structure, with the teacher reminding the learners to produce full sentences. The teacher's follow-ups concerned both content and language, e.g., the difference between descriptions and interpretations.

In the fourth lesson, 17 students were present. This lesson continued where the previous had left off, namely the comparison of the findings (task C/c). Here, all learners seemed to come up with the same point, which was that the first paragraph of the Duke's text and the portrait corresponded, but no other links between the sources were mentioned. Since everything took longer than expected, the tasks on source D as well as the final task were left out. Then, the teacher moved on to the topic of mercantilism, asking the learners to define the term without showing them the flowchart prepared. This visibly confused the learners. Then, the teacher handed out worksheet 2 which included the flowchart illustrating mercantilism, which seemed to resolve the learners' confusion. From then on, they kept to the following procedure: First they worked on one task, and then they immediately compared before moving on to the next task, etc. The answers tended to be quite proficient, and many students indeed seemed to consider the language boxes. During each working phase, the learners started working quietly and subsequently began to whisper some more. The overall working pace in this lesson was rather fast, and hardly any German was used. In the comparison phases, the teacher tended to comment the use of phrases and the language boxes, although some students seemed irritated by that. For example, NNM05 produced an appropriate sentence but did not include any of the linking devices of the box next to the task; therefore, she thought that her sentence was "wrong". In this lesson, tasks 1 to 5 of the worksheet on mercantilism were completed.

In the fifth lesson, 15 students were present. Continuing where they had left off, they started with task 6, looking at their previous answers and noting general differences between summaries in German and English. When doing the task focusing on nominalisations, some students seemed to do other things and, generally, did not appear to be interested. As a next step, the teacher asked the learners to briefly revise core concepts of mercantilism, i.e., last lesson's content. Again, this whole-group discussion followed a typical IRF pattern. Next, they moved on to the final task, i.e., working out problematic aspects of mercantilism. Again, the prompt "take notes and discuss in pairs" led to confusion. During the pair work phase, most learners talked in English and used some of the phrases (hedging, linking) too. When sharing and discussing their ideas in the plenum, it

became apparent that some learners had substantial gaps when it came to related topics (that had been or should have been covered in previous grades). Thus, the teacher spent quite some time here clarifying concepts like *triangular trade* or *colonialization*. Here, the teacher switched to German once when talking about something they should have learned in lower secondary.

### 7.3.2 Cycle 2

The materials developed in cycle 1 and improved for group B were implemented in school B over the course of five lessons in March and April 2019 (26<sup>th</sup> & 29<sup>th</sup> of March; 3<sup>rd</sup>, 5<sup>th</sup> & 10<sup>th</sup> of April).

In the first lesson, 20 participants were present (plus three non-participants excluded from the transcripts). Like in cycle 1, the unit began with an introduction to the topic using slides provided, activating prior knowledge and providing working knowledge. This part was strongly teacher-led, with students just answering a number of questions and remaining passive for the most part. The teacher used English most of the time but sometimes switched to German to clarify terminology, repeating some central explanations and addressing students who did not seem to be participating. These students often replied tersely. The contributions by students volunteering to participate, in contrast, were quite substantial and presented fluently. After roughly 25 minutes of introduction, the teacher handed out the worksheet and explained the procedure. The students then started working silently on their laptops for the remainder of the lesson. The teacher was walking around, answering questions and looking at their progress, once reminding the class to consider the language boxes.

In the second lesson, 17 participants were present (plus three non-participants). In this lesson, everybody worked on the tasks at their own pace. Once finished with one part, the learners were supposed to raise their hand to find someone of similar speed to compare and discuss answers with. In the meantime, the teacher was walking around the room, answering questions, monitoring progress, and providing feedback. The first ones completing source A compared answers after ten minutes into the lesson. Others needed substantially longer, finishing source A after 25 minutes when others were already working on source C. In any case, the comparisons and discussions were usually quite brief. Against the teacher's suggestion, some students always compared with the same person rather than moving around. Yet, some of the pairs recorded switched partners, and most of them used English consistently, made use of the phrases provided, and discussed the points more thoroughly. Overall, the lesson started rather silent but became livelier towards the end.

In the third lesson, all students were present. They were given another 20 minutes to finish the tasks. Like last time, some seemed to work in a focused way, thoroughly discussing their answer with peers and also completing the optional task (D), while others were procrastinating from time to time. Again, the teacher was walking around, monitoring their progress, answering questions, and providing feedback. He also clarified for everybody what "assessing historical significance" typically involved. After 20 minutes, he told them to stop working and pointed out that those that

were absent at one point should mention this in their file, and if they decided to finish the tasks at home, he would give them extra credit. Then he asked them to fill in a *Mentimeter* survey to create a basis for a joint discussion of the topic. The guiding questions for this online survey related to features of a good king and what we could learn from Louis XIV. Based on their answers, T<sub>B2</sub> summarized the main aspects present in the sources and linked them to the present. Some students seemed really interested and were willing to discuss, often providing rather eloquent contributions, while others remained passive or only added very short input. Mostly, this summary and the ensuing discussion were conducted in English. Only one student (USN05) asked whether to answer in English or German, which the teacher replied with “try in English”. The student then provided a relatively good answer in English.

In the fourth lesson, all students were present again. The teacher now told them to open worksheet 2 on their laptops, showing them the tasks and the flowchart via the projector. He first talked about components of a good definition, providing a subject-specific and a mundane example. Like in the other group, the students first tried to do the individual tasks on their own before comparing in class and moving on to the next task. They worked rather fast and efficiently, yet they were sometimes reluctant to share their answers with the class. Some students also seemed preoccupied with other things. A number of students, however, were quite engaged in the lessons, producing complex answers in the target language. Others started in English but switched to German when they got stuck. The teacher usually commented the content of their answers, linking them to present-day issues, but sometimes he also addressed their use of phrases. Moreover, he clarified vocabulary every now and then. After 30 minutes, task 1 to 5 were completed. Task 6, i.e., reviewing their summary and guiding their attention towards the style of the language used, led to some confusion, both on the teacher’s and the students’ part. It was then communicated that the learners should paste the individual answers to the tasks into one text and add a conclusion, read through it, and ask in case anything was unclear content-wise. No questions came up, coinciding with the end of the lesson.

In the fifth lesson, 17 participants and three non-participating students were present. The teacher did a quick revision using the app *Padlet* to check what they remembered. Most students mentioned “one ruler”, “active trade balance”, “more exports than imports”, or “state-run economy” in this brainstorming activity. Other key words were “colonialism”, “exploitation”, “dictatorship”, “torture”, “elegance”, “abolishment of domestic tariffs”, or “centralized economy”, amongst others. Some also added complete definitions like “a form of government in which all power is vested in a single ruler or other authority”. The teacher evaluated some of the answers and picked up a number of terms, starting a plenary discussion and connecting some of these ideas to current politics. Then after just 12 minutes, this unit was concluded.

### 7.3.3 Cycle 3

The second unit in school A started right after the Easter break in late April 2019 and focused on the Industrial Revolution (IR). Nine students were on time, four learners arrived late, and four more were absent. The teacher handed out the first worksheet, explained tasks 1-3, pointed out that there were boxes intended to help them, and specifically clarified the meaning of the causal cohesive devices in the language box. She set a timeframe of 20 minutes and then let them work on their own, reminding them that everybody was responsible for their own notes, but they could and should talk about their outcomes with their peers. In general, the students seemed quite focused, and about half of them did the tasks in pairs and half on their own. The students appeared to mostly use English when completing tasks together or comparing their results. They also helped each other out, explaining tasks and parts of the text but also co-constructing answers. Meanwhile, the teacher was walking around helping students, clarifying terms, and reminding them to consult the glossary. She also provided individual feedback and instruction, both in terms of content and language (e.g., concerning the use of passive voice or how to implement the phrases from the box). Sometimes she approached learners that appeared to be needing help whereas others directly asked for assistance for both content- and language-related issues. They then compared their answers in class, with most students participating actively. Most of the answers provided by the learners were quite appropriate, using nominalisation and cause-effect phrases (especially for exercise 2), some of which were provided by the boxes, as well as reporting verbs for answers to exercise 3.

The next day, 16 students were present for the second lesson of the module. This time, they were working on exercise 4, i.e., the gap-fill exercise. Again, they all seemed to be working for roughly ten minutes, some on their own and some together with a partner. Here, the construction “movement to + noun” confused them right in the first paragraph, so, once they moved on to the comparison phase, the teacher decided to start with item 2 instead. The learners seemed to enjoy the task and most answers suggested were correct. Only item 1 and 7 were quite difficult for them, potentially because they could not work out the rather formal syntax of the sentences to be matched (“movement to” + noun, nominalised construction “inventiveness”). The teacher often added comprehension questions, and sometimes she tried to connect to the previous module on mercantilism. Not many, however, seemed to remember much, and so the teacher switched to German to check what they could remember. Some then added what they remembered using German, but, in general, not much was mentioned. After finishing the comparison, the teacher zoomed in on the third paragraph and collected ideas why population grew during the IR. The students’ ideas were rather vague and not always historically valid (e.g., migration, wedding boom, religion). The teacher guided them to more relevant and appropriate ideas (e.g., better nutrition) and then let them start with part 2 of the unit, namely the graph description. The teacher first clarified the terms “rural” and “urban” and then gave them ten minutes to complete task a-c, adding that those that finished early should also complete task d. When comparing their

answers, the students often included the phrases provided, which the teacher praised explicitly. Some students, however, struggled with this exercise, especially with task b. Here, answers were co-constructed by several students, prompting words, rephrasing their peers' answers, and adding information, with some additional teacher guidance. The students also asked the teacher about the use and meaning of some of the phrases provided in the box (e.g., "remain steady"). The comparison of task c and d was done swiftly, finishing part 2 at the end of this session.

In the third lesson, 14 students were present. After briefly revising the concept of urbanisation, the teacher explained part 3 (visual sources) and ensured that everybody understood what the performative verb "describe" entailed. The students then completed the worksheet rather silently and presumably in a focused way. After just five minutes, they compared their answers and observations. In general, the students noticed many aspects. When moving on to a more interpretative task (b), the teacher asked one student to read out the grey box, reminding them to provide reasons for their views. Some used hedging devices when they speculated about the people and circumstances depicted in the picture. Some learners even overused hedging devices, which the teacher commented on. They then moved on to the impulse question dealing with potential motives for taking pictures. While discussing this question with their peers, they code-switched quite frequently and eventually came up with a substantial list of potential motives when taking pictures. Next, the learners were asked to discuss potential reasons why the artist drew the visual source on the worksheet the way he or she did. Here, the teacher reminded them to use hypothetical language when speculating (task c). Like usual, the students worked on this task individually or in pairs while the teacher was walking around, clarifying vocabulary (and prepositions), and providing feedback. When comparing ideas, most students did use hypothetical language, including modal verbs, correctly. The teacher also explicitly linked back to the impulse question and the learners, indeed, dug deeper when it came to the motives of the artist. Additionally, the teacher modelled some answers, emphasizing modal verbs, which some students wrote down on their worksheets. Moving on to the textual sources of part 3, the learners completed these tasks individually, silently, and fast. Comparison of tasks, too, did not take very long. Here, the learners employed quite a nominal style at times (e.g., "couldn't adapt to the rapid speed of growth"), and those that did not contribute anything orally took notes.

The fourth lesson took place in a computer room with 16 students present. In this lesson, the learners worked on part 4, i.e., the written report about working conditions during the IR from the perspective of a 19<sup>th</sup>-century English factory inspector with the help of the sources provided. Together, they went through the 'things to consider box', and then the students were given 30 minutes to complete the assignment. The students seemed very focused, highlighting information and typing up their text. Some students asked questions, but overall, everybody was working individually without help. As the lesson progressed, some students started talking to each other, mostly discussing their selection of sources, specific formulations, and the length of their texts. Additionally, two students who searched for their own sources on the internet realized that they



could have used the sources provided. Those that finished early printed out their text and showed it to the teacher, who provided feedback in terms of content and language, which the learners then implemented. Some students (like ORH09) who finished early even with the corrections subsequently moved on to help their peers (e.g., HIP11). After half an hour, everybody handed in their texts, and they moved on to the final task on the handout, i.e., a comparison between past and present labour and industrial production. Mostly those students that usually contribute a lot participated actively, resulting in a fast-paced discussion on a relatively high level.

In the fifth lesson of the module, with 16 students present, this discussion was resumed, first by summarising and revising the notion of industrialization before comparing historical processes to current ones. Defining the concept was still not a straightforward task for some, so the teacher had to elicit relevant aspects step by step. When comparing the past and the present, the teacher added the learners' ideas onto the whiteboard, juxtaposing the IR and today's industrial production. Here, many students contributed; stronger students added more to the discussion and in full sentences, whereas weaker students were less involved and often contributed just phrases or incomplete clauses. In the beginning, there was a bit of confusion what exactly should be compared, but then the discussion became quite lively, including aspects about current politics (e.g., 12-hour-working days in Austria, digitalization of teaching and learning, or migration). The teacher also tried to include students that remained quiet. If these learners then added a superficial or vague answer, the teacher tried to elicit more. Once they switched to German for this purpose, but eventually the student (NNM05) simply gave up. After less than 20 minutes, they concluded this discussion and the unit on the IR.

## 7.4 Evaluation of interventions

In this subchapter of the analysis part, the outcomes of each cycle's evaluative processes are described and examined. To allow a comprehensible but detailed analysis of the individual evaluations of each cycle, this subchapter is, first of all, split into the three cycles. Then, each cycle is divided into interviews and post-intervention written tasks.

Concerning the interviews, corresponding analyses of students and teacher interview are combined in one section since learners and teacher talked about the same aspects in their respective interview. Additionally, whenever appropriate and insightful, the teacher's or students' perceptions are corroborated with episodes taken from the lesson transcripts. The interview subsections are further divided into (a) *experiences with the intervention*, (b) *evaluation of the intervention*, addressing various facets from level of engagement to the educational value of tasks or their satisfaction with the language support, and (c) *summary and implications for the design* to be considered in upcoming design sessions. It should be mentioned that sometimes, participants provided further information on the context of the study, including learner needs or background information on the participants or their school, and the teachers also shared their reflections on the process and the outcomes of the study. To keep the sections focused, these points are not

presented in a distinct subsection (unlike other main categories of the qualitative analyses) but included elsewhere whenever relevant.

Turning to the post-intervention written tasks, after a brief general overview of the results, the subsections are subdivided into (a) *history-based rating*, (b) *linguistic rating*, and (c) *summary and implications for the design*, similar to the structure of the needs analysis subchapter. In this subsection, the post-intervention results are qualitatively and quantitatively compared to their pre-intervention counterparts, thereby illuminating to what extent the learners' performance changed throughout a cycle.

Content-wise, this subchapter sheds light on all three RQs:

- RQ1: What kind of content-and-language-integrative pedagogical measures and materials (type and features) are needed to help students improve and elaborate their verbalization of cognitive processes (CDF use) as
  - a) perceived by learners
  - b) reported by teachers
  - c) observed in written student performances?
- RQ2: How do students respond to explicit teaching of CDFs in the history CLIL classroom as
  - a) reported by learners
  - b) perceived by teachers
- RQ3: What is the effect of CDF-oriented teaching on the learners' development of historical competences and academic language skills as observed in written performances?

The interview data mainly provides insights in relation to RQ2. Here, it should be noted that the retrospective interviews with the teachers only offer insights from an external perspective, while the students offer an internal perspective of their experiences. As such, the analysis of subthemes usually starts with the perspective of the learners, which is then compared to and complemented with their teacher's view on the respective topic. RQ3, then, is based on the learners' written performances. Finally, all these insights indirectly feed into the empirical base of RQ1, which is concerned with the type and features of effective content-and-language-integrative materials, creating a basis for improving and adapting the intervention and, on a wider scale, our approach. As with the previous sections, the focus will lie solely on the empirical base. Only in subsections c), i.e., the summary, (some) links to relevant literature are provided. The full discussion of the findings follows in chapter 8.

In [section II of the appendix repository \(data analysis\)](#), one can access analytical steps and products to allow a transparent and comprehensible analysis. This part of the appendix contains summary tables, MaxQDA (2020) code-matrices, code co-occurrence models, and hierarchical codes/sub-codes models with frequencies of all interviews ([subsection A – interviews](#)). Concerning the post-intervention tasks, SPSS calculations (descriptive statistics including tests of normality, correlation analyses, and tests of comparison including effect size calculations) can be viewed ([subsection B – pre- and post-intervention tasks](#)). Rating spreadsheets as well as original data and their qualitative analyses (MaxQDA files) are only available on personal request to make sure that the data is only used for authorized research purposes.

## 7.4.1 Cycle 1: absolutism and mercantilism (school A)

### 7.4.1.1 Retrospective interviews with students and teacher

The retrospective interview with five students of group A (two male, three female) took place in early March 2019 right after the completion of the intervention. As not all students who participated in the needs analysis interview were present on this day, the composition of the groups differed slightly.<sup>91</sup> The retrospective interview with T<sub>A</sub> was conducted one month later.

#### a) Experiences with the intervention

To begin with, both teacher and students experienced the unit on absolutism and mercantilism as very different to their normal history lessons for a number of reasons. First of all, the learners were not used to such a level of labour-intensity, including a high amount of writing and the use of step-by-step worksheets, ultimately exhausting some of the students. Normally, the students reported that they would 'just talk openly' about the topic in the form of whole class discussions (in which only three or four people would participate, according to the learners). The teacher's account corresponds to this description, stating that usually, she would neither structure her lessons in such small steps nor work with historical sources that much. With the lessons being denser, another issue surfaced during the first cycle, namely that the students' work paces differed extensively. Another major difference repeatedly mentioned in both interviews was the considerable focus on linguistic expression in the subject history. Here, T<sub>A</sub> reported that she struggled with finding the right extent of focusing on language when comparing results. During whole-class instruction, T<sub>A</sub> often focused on the phrases included in the linguistic support accompanying the tasks by telling the learners to rephrase, correcting their use of phrases, or providing recasts, but also by asking for different phrases rather than for different ideas when comparing answers. In other words, in whole-class teaching, linguistic form was often prioritized over content, as can be seen in the following extract:

#### Extract 77, lesson 2, cycle 1:

- 1 TA: Who would like to start reading out his or her answers? [...]
- 2 ORH09: First of all, a good king should be guided by his naturally good sense and moreover his work should be agreeable and ähm, not only should be well, should he be well informed about his own state but also about foreign politics.
- 3 TA: Mhm. Anything to add?  
((EVA02 raises hand, TA nods in his direction))
- 4 EVA02: And he should have general, good general knowledge of the world?
- 5 TA: Äh, okay. **Could you use one of the expressions** in the box please?
- 6 EVA02: **Additionally**, he should have good general knowledge.
- 7 TA: Mhm. [affirmative] Anything else? ((HIP11 raises hand))
- 8 TA: HIP11?
- 9 HIP11: And additionally-
- 10 TA: = **Na, another one please. Another phrase.**

---

<sup>91</sup> Participants in the pre-intervention interview: ARJ01, HIP11, IKS12, OPB04, SAA03  
Participants in the post-intervention interview: ARJ01, ETS12, HIP11, OPB04, ORH09

In the case above, the teacher does not seem to be really interested in the content the learners want to deliver, i.e., features of a good king according to the text, but the phrases they use for introducing their ideas, as she only comments on their linguistic choices. As similar episodes happened throughout the unit, at some point, learners felt like their answers were wrong if they did not use as many phrases as possible:

**Extract 78, lesson 4, cycle 1:**

- 1 TA: Which resulted or led to cheap production. Good. NNM05?
- 2 NNM05: Frau Professor, nein. [No, Ms professor]
- 3 Sfx: Lies einfach vor. [Just read it out.]
- 4 NNM05: [to students] **Nein, es passt nicht.** [No, it's not right.]
- 5 TA: Come on.
- 6 **NNM05: First of all, cheap production was a result of lower transport costs, lower wages and low price of corn and cattle.**
- 7 TA: Good. And now explain to me why it's wrong. I mean you said before, das passt nicht. Warum nicht? [that it isn't right. Why not?]
- 8 NNM05: Ja, **weil ich die zweite Spal-, also die zweite Phrase nicht einbaut hab.** [Yeah, because I didn't include the second col-, well the second phrase.]
- 9 TA: Okay. Yeah, but the sentence is perfect. Okay?

In turn 6, NNM05, a student with rather low levels of achievement in history and English as well as active participation in the units observed, presented a perfectly correct answer, both in terms of content and language. Yet, she did not want to share it with the class since she had not included the second language tip accompanying the task. This episode is a good example of how too much focus on form and explicit treatment of language support in content classes might discourage learners that already struggle with motivation. These sentiments also transpired in the focus group interview, as it was mentioned repeatedly that the use of certain phrases should not be the criterion for discussing content:

79	English translation	Original quote
OPB04 & ORH09	<p>OPB04: Yes, and <b>I didn't like it that much that the words were so prescribed</b> [...] instead of just writing what one was thinking, one had to check whether they [= the phrases] were included. [...]</p> <p>ORH09: At this point, we are, I think, well many of us, good enough to have our own way of expression and if you are told how the words should be linked, writing gets complicated somehow, yeah.</p> <p>OPB04: <b>Yeah, and even if your sentence was correct, you had to do it all over again because the phrases were not in there.</b></p>	<p>OPB04: Ja, und ich mocht es auch nicht so, dass die Wörter so vorgeschrieben waren [...] statt das zu schreiben, was man denkt oder so, man musste schauen, ob man das eh drinnen hat [...]</p> <p>ORH09: Mittlerweile sind wir glaub-, also viele sind gut genug, dass sie eine eigene Ausdrucksweise haben auf Englisch und wenn man dann genau gesagt bekommt, wie man die Wörter verbinden muss, dann ist es urkompliziert so zu schreiben irgendwie ja.</p> <p>OPB04: Ja und auch wenn der Satz so richtig war und man ihn dann einfach nochmal komplett neu machen musste, weil man das jetzt nicht drinnen hatte.</p>

Although students do appreciate being corrected when making a mistake (as they argued in their initial interview, see p. 149), they were annoyed that during the lessons, linguistic form was prioritized over the actual content they wanted to convey. In the teacher interview, T<sub>A</sub> argued that she felt that she had to enforce the learners' use of these phrases to ensure uptake:

80	English translation	Original quote
T <sub>A</sub>	<p>TA: But <b>I felt that I have to do this</b> [= enforcing the phrases] <b>to make it stick.</b></p> <p>R: Mhm [agreeing], I see. Well, maybe this is something that one should do every now and then to some extent, ideally.</p> <p>TA: Yes, yes.</p> <p>R: But of course, it's difficult now with such a project, which is rather dense.</p> <p>TA: = yes, very concentrated.</p>	<p>TA: Aber ich hab das Gefühl gehabt, ich muss das tun, damit, das hängen bleibt.</p> <p>R: Mhm [zustimmend], na eh. Also, das ist wahrscheinlich was, was man über längere Zeit immer mal wieder wenig am besten.</p> <p>TA: Ja, ja.</p> <p>R: Es ist natürlich schwierig mit so einer, jetzt mit so einem Projekt, das was da so, ähm, gedrängt ist</p> <p>TA: = geballt ist, ja.</p>

We agreed that focusing on phrases and enforcing their use is more beneficial if done sparsely but regularly to avoid student frustration. Yet the organisation of the project required a certain level of density in this regard. In general, however, the teacher felt that explicit attention to language was necessary:

81	English translation	Original quote
T <sub>A</sub>	<p>I do that anyways, well full sentences [...] because otherwise I have to interpret 'okay what does he mean?'. And I don't want that, and anyways, <b>it's important for the final exam and everything.</b></p>	<p>das tu ich sowieso generell im-, also ganze Sätze [...] weil dann muss ich interpretieren, „okay was meint er jetzt damit?“. Das will ich gar nicht, außerdem ist das ganz wichtig auch für die Matura und für alles, also.</p>

With these considerations in mind, it was agreed that insisting on full sentences and the use of certain phrases some of the time was crucial for the final exam as well as success on a more general level (“everything”), but it should not be the main criterion when talking about content.

When looking back to the pilot lessons as well as previous interviews, it appears that T<sub>A</sub>'s beliefs concerning whether to include linguistic aspects in her content teaching have changed considerably. In the pilot lessons, the phrases were not enforced and hardly any comments on linguistic features could be observed. What is more, in the initial interview, she stated that she did not teach language in her history lessons but was unsure how to continue, while the students of group A felt that T<sub>A</sub> had already considered linguistic aspects prior to the project. After the first unit, it seems that T<sub>A</sub>'s insecurity in this regard led to overcompensation, resulting in student frustration as could be observed in the lessons and was also put forward in the retrospective interview with the learners.

Figure 21 on the next page summarizes the participants' experience by showing which aspects of the implementation the students and teacher perceived as positive or negative. It seems that the linguistic pedagogical practices implemented when teaching this unit negatively affected the participants' experience, as repeatedly put forward by the students. The teacher, who only talked about this briefly, described her experiences as mainly positive since she felt that the scaffolding really supported the learning process.

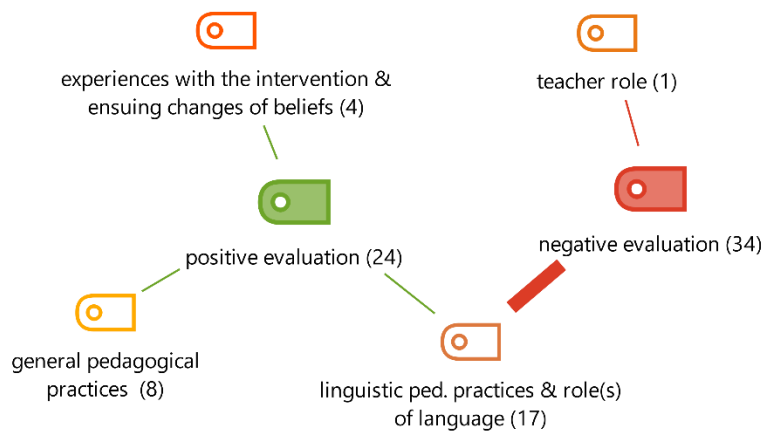


Figure 21. Code co-occurrence model: experience of students and teacher, group A, cycle 1

### b) Evaluation of the intervention

The code-relations models below (Figure 22 & Figure 23) visualize which aspects the students and the teacher evaluated in their interviews, how often they mentioned these points, and whether they assessed them positively, negatively, or inconclusively. On first glance, one can see that the students focused a lot more on what they perceived as negative (13 positive, 24 negative, and 2 inconclusive markings), while the teacher discussed positive and negative aspects almost to an equal extent (11 positive, 10 negative, and 5 inconclusive markings). For most sub-codes, positive and negative aspects were mentioned in both interviews.

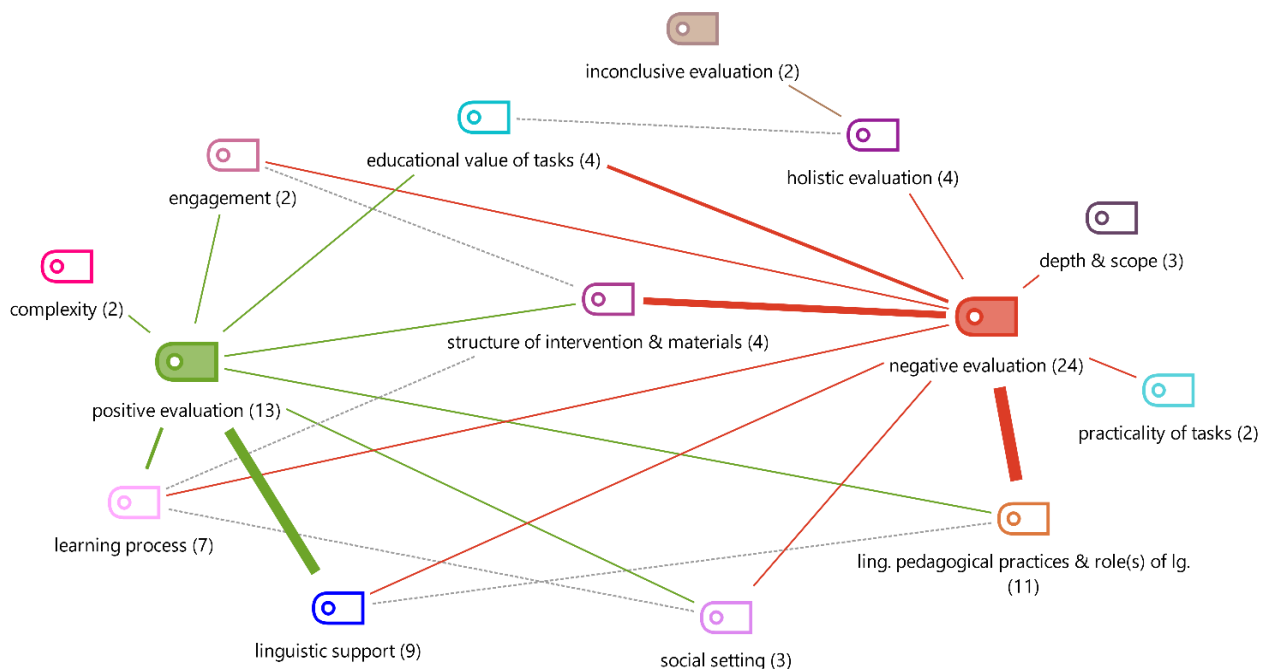


Figure 22. Code co-occurrence model: evaluation by the students of group A, cycle 1<sup>92</sup>

<sup>92</sup> In these code co-occurrence models, some labels of the codes were slightly altered to better represent their definition (e.g., the super-category *evaluation of intervention* was relabelled into *holistic evaluation*) or shortened to save space (e.g., ~~comments on~~ *linguistic support*).

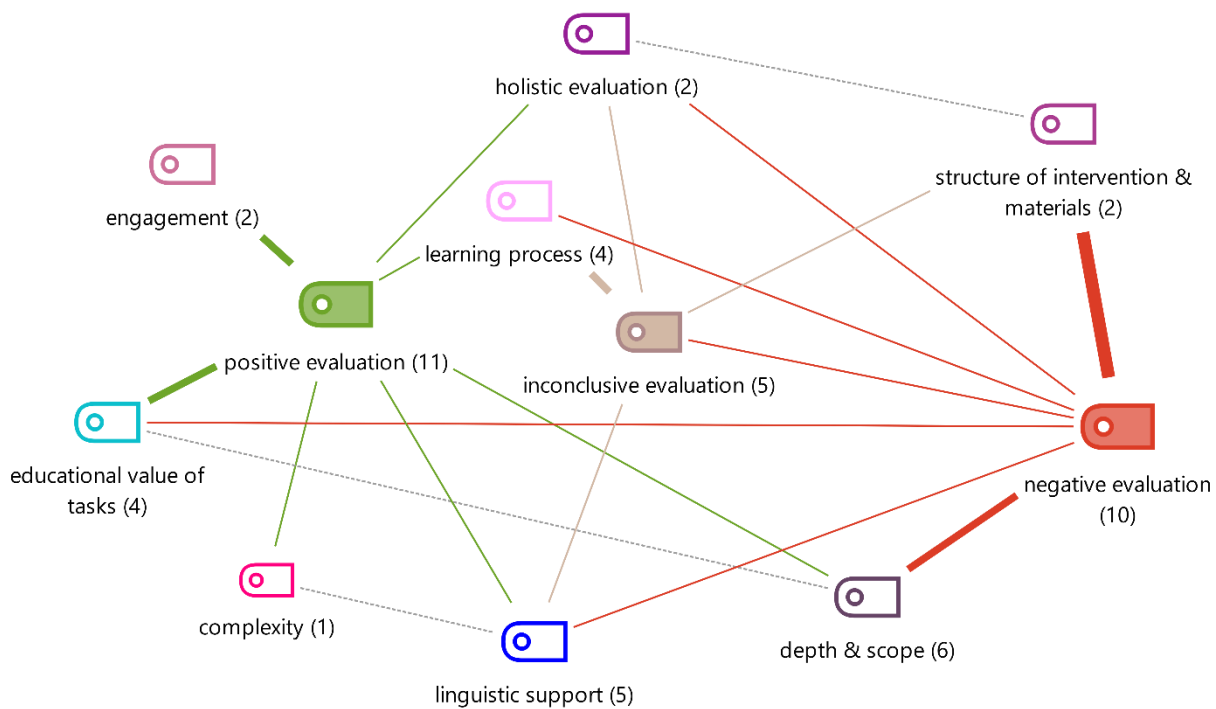


Figure 23. Code co-occurrence model: evaluation by TA, cycle 1

The level of **engagement** was mainly assessed as positive since both students and teacher reported that learners were more active than usual, including learners who usually do not participate much:

82	English translation	Original quote
TA	Well, <b>they have all worked really well and dutifully.</b>	TA: Also sie haben wirklich alle gut, alle brav gearbeitet
ORH09 OPB04 Sfxx	ORH09: There, well, there <b>were more people participating</b> , I guess, than normal. OPB04: Yes. R: And do you think this was the case because you were being filmed or because the task required it? Sfxx: <b>Because the task required it</b> , yes. [...] ORH09: Because usually there are three, four people, maybe, participating and now we were practically forced to participate because everybody had to work it out themselves and was not prompted what to say or write.	ORH09: Es, also es haben irgendwie auch mehr Leute mitgearbeitet, hab ich das Gefühl, als sonst. OPB04: Ja. R: Und glaubt's ihr, dass das deshalb war, weil ihr aufgenommen wurdet oder weil's die Aufgabe erfordert hat? Sfxx: Weil die Aufgabe erfordert hat, ja. [...] ORH09: Weil es gibt halt immer so, sonst halt drei vier Leute, die vielleicht mitarbeiten und jetzt wurden wir quasi gezwungen mitzuarbeiten, weil hat jeder für sich ausgearbeitet hat und nicht schon vorgesagt bekommen haben, was sie zu sagen, oder zu schreiben haben.

Later, the students added that they did not participate as much when doing tasks that only asked them to "discuss with a partner" because they did not feel as responsible, which is something the teacher seemed to confirm. This point is related to the **social setting** of these tasks. In the student interview, it was mentioned that in this class, pair work that was based solely on discussing content without requiring both learners to produce any form of written output was problematic due to the students' tendency to skip these discussions. Instead, the students prefer "individual

work”, so that they feel responsible for their own progress. The students explained, however, that by “individual work” they meant being forced to take notes themselves while still completing the task in pairs, nonetheless. In other words, these students did not ask for a different social setting after all but for a different way of framing pair work, i.e., providing prompts that stress their own accountability when it comes to taking notes and recording output.

Turning to the level of **complexity**, the teacher and the students present felt that the level of complexity was appropriate. Yet, the students suspected that the materials might have been too challenging for some of their peers struggling with English and especially text comprehension as well as with history or motivation for school in general. But they also added rather cynically that, for some, everything would be either too complicated or too boring and that they could think of nothing that would improve their colleagues’ learning progress apart from “forcing” them to participate, which is something the students attributed to the tasks of this unit, as mentioned above.

Regarding the **learning process**, the biggest issue revolved around the different paces of the learners when completing the tasks. The students complained repeatedly that they were annoyed by the frequent waiting times. According to the teacher, this class was very heterogeneous in terms of pace and achievement level. For the teacher, this aspect was less of an issue with the worksheet on mercantilism as it was well scaffolded, including many small steps and frequent whole-class comparisons, keeping everybody cognitively involved. Another aspect pointed out by students and confirmed by the teacher was that some parts of the source-based tasks (i.e., worksheet on Louis XIV) appeared somewhat repetitive, which again affected the more proficient learners negatively since they would have welcomed more different challenges. So, in terms of **overall structure**, some sequences of tasks did not provide enough variety in terms of content or task-type, and overall, the worksheet took too much time. Once finished with the tasks concerning source C, the learners had already grown tired of doing source analyses on the same historical figure so that the teacher decided to skip source D, also to be able to keep to the schedule. For the stronger students in this group, the steps of the tasks were too small and the whole structure too compartmentalized. They seemed to favour a more holistic approach combining various tasks to allow for a more open discussion of the topic. They further argued that, although they liked “summarizing” the source before actually talking about it, they felt that this preparation step should be less guided and more open to account for individual differences:

83	English translation	Original quote
ORH09 ARJ01	<p>ORH09: <b>Summarizing, preparing on your own, I do find really good</b> (.) because for a lot of our classmates (.) but splitting into these questions, that was quite unnecessary for me and simply a waste of time, to be honest.</p> <p>ARJ01: Yes, one could have summed up the first page in one question and the second page in one question.</p>	<p>ORH09: <i>Das mit dem zusammenfassen, selber ausarbeiten, find ich eigentlich eh urgut (.) weil für viele aus unserer Klasse (.) aber halt so, dass man so auf Fragen aufgeteilt hat, das ist für mich halt irgendwie unnötig und einfach nur Zeitverschwendung, to be honest.</i></p> <p>ARJ01: <i>Ja man hätte, die erste, die erste Seite als eine Frage machen können und die zweite Seite als eine Frage. [...]</i></p>



	ORH09: Or <b>everybody just summarizes however it works best for oneself</b> . It doesn't have to work for my teacher, it has to work for me so that I can study well.	ORH09: Oder einfach jeder fasst es so zusammen wies für sich besser ist. Muss ja nicht meiner Lehrerin passen, es muss ja mir passen, so dass ich's gut lernen kann.
--	--	--

What is interesting here, too, is the students' use of "summarizing". For them, working through a source seems to be a process of gathering and structuring information provided by a source, when, in fact, it should be a more analytical and critical process, evaluating the source and the validity of the content it presents. Thus, it can be argued that providing guidance as to which kind of steps are indicated appears to be absolutely necessary at this stage since the learners were not yet sufficiently familiar with the methodological script of source analysis. This view is shared by T<sub>A</sub>, who appreciated the clear structure and the small steps, helping the students understand how to approach sources more systematically. Additionally, the students called for a less rigid structure and fewer steps even though they seemed to be aware that smaller steps could somewhat counteract different paces and long waiting times.

Relating to the considerable differences in learning pace in this group, the teacher felt that our lesson plan was too ambitious in terms of **depth and scope** for the time planned. The students, too, argued that there were some tasks one should exclude, as the students in this interview perceived them as unnecessary:

84	English translation	Original quote
ORH09 ETS12 ARJ01	ORH09: Yes, yes that one [= nominalisation task] was extremely unnecessary. ETS12: = <b>It didn't fit in at all. For me, it felt like having English class instead of history, to be honest.</b> ORH09: Yes, and I think we all know what a noun is by now, now in third grade HAK, it would be really sad if not everybody ARJ01: No, not everybody, not everybody.	ORH09: Ja, ja das [= nominalisation task] war extrem unnötig. ETS12: = Das passt gar nicht rein-. Es kam mir irgendwie vor als hätten wir Englisch statt Geschichte ehrlich gesagt. ORH09: Ja und also ich glaub wir alle wissen was ein Nomen ist, jetzt in der dritten HAK, also wär sad wenn nicht ARJ01: Nein, nicht alle. Nicht alle.
ETS12	<b>I found the last page unnecessary for the most part because we had already talked about it</b> , his features, what he's like, and so on.	Ich fand irgendwie die letzte Seite meistens unnötig, weil wir davor schon gesprochen haben, seine Eigenschaften, wie er so ist und so weiter.

By calling tasks "unnecessary", the students implied a low **educational value** for history classes, either because these tasks "felt like English class", were below their level, or repetitive in terms of content. However, not all students agreed on all of these points. Similarly, the teacher did not agree with all cuts suggested by the learners. For example, while the tasks for source A and B were appreciated by the students, source C was rejected by some of the learners on the grounds that an analysis of a painting would be too obvious, thus unnecessary. Still, T<sub>A</sub> would keep the source and two out of three tasks, namely those that address the learners' issue with confounding describing and interpreting (task a and b) and their lack of awareness concerning the methodological steps involved in source analysis. Furthermore, T<sub>A</sub> would also recommend keeping most tasks dealing with the flowchart on mercantilism, while some of the students argued for cutting or combining

some of the tasks of this worksheet. The only activity both parties agreed to cut or modify was the nominalisation task since this was too disconnected from the content.

In terms of **linguistic support**, the teacher agreed with the students that the boxes dealing with historical literacy, i.e., those that illuminate the relationship between linguistic choices and the subject history, were helpful. Furthermore, T<sub>A</sub> also appreciated the boxes that inform about the communicative intention underlying the performative verbs used in the tasks. As for the boxes offering phrases for realizing the different language functions, the students indeed thought that they could help them in their output, but they strongly argued that the use of these phrases should be voluntary and not at the centre of the lesson, as already mentioned above and visible in the following extract:

85	English translation	Original quote
HIP11	Well, I, for example, struggle in English. On the one hand, <b>it's good that the linking devices were provided, but I don't like that you had to use them by all means.</b>	<i>Also, ich hab zum Beispiel Probleme in Englisch. Einerseits ist's eben gut, dass die Linkers vorgegeben waren, aber ich find's nicht gut, dass man sie unbedingt benutzen musste</i>

In the interview with T<sub>A</sub>, we decided to keep these phrase boxes on the worksheet, but in whole-group discussions, the boxes should not get a great deal of attention. Instead, the linguistic support should be dealt with more subtly and when necessary. Concerning type of linguistic support, ORH09, a rather proficient student, argued that the support measures should not include basic or general tips for bilingual students, even though she seemed to be aware that some students really struggled with these aspects:

86	English translation	Original quote
ORH09	But I find that one should kind of expect that, especially in a bilingual HAK [...] <b>I know that it doesn't work [for everybody], but I find, well, then you can also add something like 'please don't forget to breathe'.</b> ((laughs))	<i>Aber halt ich find, dass man das quasi voraussetzen sollte, vor allem in einer bilingualen HAK [...] ich weiß, dass es nicht funktioniert, aber halt an sich find ich's, also dann kann man auch sowas dazu schreiben, bitte halt das Atmen nicht vergessen ((lacht)).</i>

Thus, the teacher and I agreed to make sure that the language tips should not be too general but tailored to the subject history whenever possible to increase acceptance and face validity.

In the student interview, the learners were asked to go through the materials again and use either green or red markers to highlight what they liked (green) or did not like (red). In the teacher interview, T<sub>A</sub> had a look at these markings, offering her views on the students' assessment. There were some cases where students marked one column of phrases in green and the other one in red. Here, T<sub>A</sub> offered a potential and convincing explanation: The ones the learners already knew were marked in red because their inclusion was perceived to be unnecessary or even patronizing, whereas new ones, or less familiar ones, were appreciated, as they entailed more educational value. Finally, the teacher argued that for linguistic aspects that did not appear crucial for historical literacy at this stage, such as nominalisation, pointers for those interested might be more expedient than language-focused exercises for everyone. In terms of length and positioning

of linguistic support, one student with a high level of achievement argued for having all linguistic support on an extra page. T<sub>A</sub> did not agree with her because this might overwhelm weaker learners, especially if the boxes present new or unfamiliar information, which is something that the results of the pilot study also revealed. In this regard, the teacher had the idea of letting the learners copy the language boxes on one or two pages after they dealt with them in the specific contexts as homework to solidify the input and make the material more practical for later use.

### c) Summary and implications for the design

To summarize, the teacher's experiences with the interventions were mainly positive and she intends to use the materials again but in a less dense and language-intense way:

87	English translation	Original quote
T <sub>A</sub>	<p>R: Do you intend to use at least parts of it [= the unit] again?</p> <p>TA: <b>Definitely!</b> Well, I don't know which ones, I haven't thought about that, but yes, <u>yes</u>, of course.</p> <p><b>R: So, probably not in one piece?</b></p> <p><b>TA: Exactly and also less language intense.</b></p>	<p>R: Und hast du vor, dass du zumindest Teile davon wiederverwendest?</p> <p>TA: <u>Auf alle Fälle</u>, mhm. Also das weiß ich jetzt noch nicht welche, das hab ich mir jetzt noch nicht überlegt, aber ja, <u>ja</u>, natürlich.</p> <p>R: Also nicht mehr so am Stück wahrscheinlich?</p> <p>TA: Genau und auch nicht so language intense.</p>

This also reflects the learners' main points of criticism. Overall, the students were quite critical about the intervention and did not describe their experience as predominately positive. This stands in contrast with another quite similar intervention study by Lo and Jeong's (2018), where both teacher and students received the intervention positively. In their study, the intervention consisted of a genre-based approach for two grade-7 history classes in Hong Kong. Of course, perceptions are very subjective and are thus difficult to compare, yet one could speculate that the enthusiasm of the Hong Kong students might be explained by their younger age as well as their cultural background. Still, it is noteworthy that group dissatisfaction in the present study seems to have more to do with the implementation than with the materials per se. While the learners of group A appreciated most of the language boxes and the high level of engagement brought about by the tasks, they were also frustrated by the way these linguistic tips were highlighted in the lessons, the amount of work, and the differing learning paces within the group. This tells us further that for these learners, a form-focused and/or a writing-intense approach, as has been employed in other intervention studies (e.g., Tedick & Young, 2018 or Whittaker & García Parejo, 2018), might not be accepted.

Similar to the students, T<sub>A</sub> appreciated the student-centredness, leading to high levels of engagement and student participation. Unlike the learners, T<sub>A</sub> especially liked the tasks that were scaffolded in really small steps, such as the worksheet on mercantilism, as she felt that learners could grasp the idea much better compared to how she usually taught it. Concerning the linguistic support, both students and teacher welcomed the boxes focusing on historical literacy, while only T<sub>A</sub> highlighted the boxes focusing on performative verbs. To increase acceptance by learners, it was agreed that some of these boxes could be even more history-specific. When it comes to the

boxes with the phrases, this cycle has shown that an appropriate degree of active focus on the use of these phrases is difficult to determine but appears to be crucial for learner satisfaction. In the pilot cycle, where  $T_A$  did not explicitly discuss the use of the phrases provided, students seemed very satisfied with the intervention. At the same time, considering that the pre- and post-intervention written task results in terms of linking changed only slightly, some learning opportunities might have been missed in the pilot cycle (see Bauer-Marschallinger, 2019). In this first main cycle,  $T_A$  seemed to prioritize the use of phrases over the content the learners wanted to share when discussing and comparing answers. This was met with rather negative reactions by the learners, which, in turn, might pose an obstacle to a productive working environment. Thus, it was agreed to include explicit focus on linguistic choices more subtly and always in relation to the content the learners want to express. In the end, teachers using these materials need to flexibly adapt and probably continuously adjust the intensity of active language focus to their students' level of amenability in this respect.

Other aspects in need of improvement, as pointed out by both learners and teacher, were length and overall structure, including strategies to mitigate the issue of varying paces. Apart from the recommendation of cutting a few tasks (e.g., C/c or mercantilism/ task 6), no concrete solutions were discussed at this point but were later developed in the design sessions (see subchapter 7.2). Furthermore, the students suggested including more individual work with the possibility to work on it in pairs to ensure that everybody feels responsible for their notes. At the same time, they asked for fewer handouts, less writing, and more whole-class teaching, reflecting their traditional way of teaching. As expected, it appears that, on the whole, this approach was new to both learners and teacher and still needed more refinement and presumably more exposure to work for both parties.

#### 7.4.1.2 Post-intervention written tasks

One week after finishing the unit on absolutism and mercantilism, the students completed their second competency-based written task (T2), which featured a caricature dealing with mercantilism (and absolutism to some extent) (see [appendix section I/ B/ file 1 – all written task prompts](#)). Figure 24 illustrates average developments of all students participating in all written tasks ( $n = 16$ ) in terms of content and language from pre- (T1) to post-intervention (T2).

To start with, both areas show an upward tendency, from  $M_{T1} = 1.83$  to  $M_{T2} = 2.04$  in terms of content and from  $M_{T1} = 1.37$  to  $M_{T2} = 1.83$  with regards to language. This means that, on average, these learners benefitted both in terms of content and language learning,

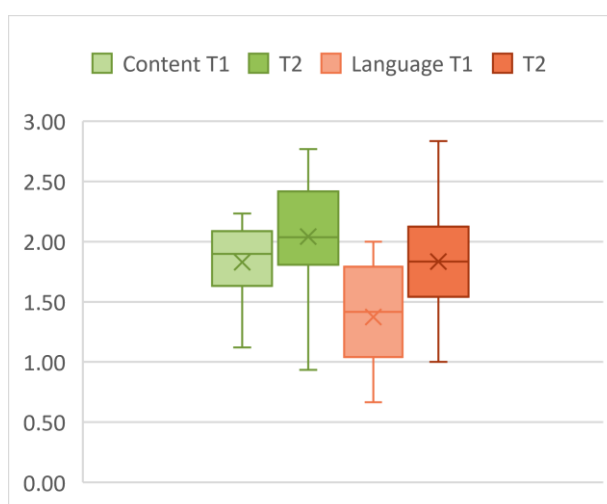


Figure 24. Comparison T1 vs. T2: boxplots, group A

with the linguistic outcomes increasing more than the content outcomes. These changes are statistically significant as measured by a paired-samples *t*-test, with  $t(15) = 2.27$  and  $p = .038$  for content results and  $t(15) = 2.86$  and  $p = .012$  for language results, presenting large effects of  $d_{content} = 2.06$  and  $d_{language} = 1.00$  (see [appendix section II/B/1/ file 4 – tests of comparison](#)).

Figure 24 further shows that the range increased, with standard deviations rising from  $SD_{T1} = 0.31$  to  $SD_{T2} = 0.51$  in terms of content and from  $SD_{T1} = 0.43$  to  $SD_{T2} = 0.44$  regarding language. The graph also demonstrates that the maximum points in both areas are considerably higher in the post-intervention sets. This could imply that for some students, these changes could be rather a result of changes in awareness than changes in skills, as the intervention only lasted for five lessons. Interestingly, those students with the highest values for either content or language in the post-task tended to be somewhat close to the group average in the first round. For example, SAA03, who achieved the maximum on the content scale (= 2.76) after the intervention, was exactly on average in the first sitting ( $M = 1.83$ ), as can be seen in Table 10. Her linguistic development is somewhat analogous, with an increase from 1.83 (= 0.45 above average) to a language mean of 2.33, which is the second highest score in the set. The maximum in terms of language was 2.83 and was achieved by EVS03, who scored only 1.20 (= 0.18 below average) before the intervention. In terms of content, EVS03 improved from 1.96 to 2.56, which takes up second place after SAA03.

Table 10. Individual results: group A, cycle 1

On the other half of the spectrum, those learners with language results of  $\leq 1.0$  in the first sitting (NNM05, OPB04, ELF03, HIP11) now all achieved 1.83 for language, which corresponds to the mean of the second sitting. However, two of these students, namely NNM05 and HIP11, performed more poorly in terms of content in the post-intervention task, together with two other students (ELF03, ATC04) whose content results decreased as well. A potential explanation could be that focusing on language might have been too much for them to process, negatively affecting their performance. Additionally, the linguistic performance of four

student code	content			language		
	T1	T2	T2-T1	T1	T2	T2-T1
ATC04	2.20	1.66	-0.54	1.50	1.33	-0.17
AVS07	1.73	1.83	0.10	1.17	1.83	0.67
ELF03	1.41	0.93	-0.48	0.80	1.83	1.03
ELH01	1.72	2.03	.031	1.17	1.33	0.17
ETS12	1.61	1.80	0.19	2.00	1.00	-1.00
EVA02	2.13	2.43	0.30	1.83	2.17	0.33
EVS03	1.96	2.59	0.62	1.20	2.83	1.63
HIP11	2.10	2.00	-0.10	1.00	1.83	0.83
ICM01	2.03	2.37	0.33	2.00	1.67	-0.33
IJT12	1.70	2.04	0.33	1.33	1.83	0.50
IKS12	1.96	2.37	0.40	1.50	1.50	0.00
LES02	2.00	2.53	0.53	1.67	2.17	0.50
NNM05	1.12	1.10	-0.02	0.67	1.83	1.17
OPB04	1.48	1.87	0.39	0.80	1.83	1.03
ORH09	2.23	2.33	0.10	1.50	2.00	0.50
SAA03	1.83	2.77	0.93	1.83	2.33	0.50
average	1.83	2.04	0.21	1.37	1.83	0.46

students (ATC04, ELH01, ICM01, ETS12) declined too. One can only speculate, but they might have been overwhelmed with the linguistic input. As for ETS12, who shows the highest decrease and the lowest score of the set, the reason could be on the affective level, as she repeatedly argued in the interview that she found the linguistic support redundant. But of course, any of these changes, positive and negative alike, might stem from completely different factors, such as topic of the task, their condition on a particular day, or their current motivation.

#### a) History-based rating

Looking at the differences between T1 and T2 in terms of history results, the most obvious result is that the average outcomes of all descriptors increased (see Table 11).

Table 11. History-based rating: T1 vs. T2, group A, cycle 1

	means			t-tests				Wilcoxon signed-rank tests				
	T1	T2	T2-T1	t	df	p*	d	T <sup>93</sup>	p*	r <sub>w</sub>	Mdn T1 T2	
target level	2.22	2.40	0.19	1.89	15	.079	0.62					
accuracy o. c.	1.74	2.03	0.29					77.0	.124	.27	1.74	2.10
systematicity	1.74	2.00	0.26	1.71	15	.107	0.81					
target comp.	2.25	2.28	0.04					54.5	.898	.02	2.40	2.45
justification/c.	1.50	1.75	0.25	2.04	15	.059	0.87					
scope o. c.	1.56	1.78	0.22	1.43	15	.173	0.76					
	*α < .05											

However, none of these differences were statistically significant as measured with *t*-tests for all normally distributed data (both T1 and T2) and nonparametric Wilcoxon signed-rank tests for the remaining data. Nonetheless, the data still shows a positive trend, as the results for all descriptors improved. Moreover, some of these findings present a large (*systematicity*, *justification/comprehensibility*, *scope of content*), medium (*target level*), or small effect size (*accuracy of content*).

The most substantial increase concerns *accuracy/relevance of content*, which could be explained with the amount of time spent on the topic. Usually, the teacher does not spend five lessons on the same topic, which also holds true for the topic of Early High Civilizations, which was featured in the pre-intervention task. Nonetheless, one could argue that the intervention was designed in a way that allowed students to engage with historical sources on a certain topic in a largely historically accurate and relevant way. *Accuracy/relevance* strongly correlated with *justification/comprehensibility* ( $\tau_b = .88, p < .001, n = 16$ ); a descriptor that also presents a considerable growth from T1 to T2 (+.25). In other words, those students who did not contradict established historical facts and selected facts relevant to the task often managed to produce a justified and comprehensible answer. In the needs analysis, one frequent issue was that the learners did not engage directly with the source, often not justifying their claims with what they saw in the picture

<sup>93</sup> Due to the small sample size, test statistic *T* is reported rather than the standardized test statistic *z*. For *z*-scores, see appendix, [section II/ B/ folder 1 \(cycle 1\)](#).

and/or what they remembered about the historical context. This has become somewhat less of an issue in the post-intervention productions. For example, when asked to assess to what extent a certain concept was connected to the source provided, the quality of EVS03's answers in pre- and post-intervention task differed quite notably:

88	DS: {In the last part of the picture, it looks like the workers are making place where the water can go through.}
Pre-task	
Item 3 EVS03 (A)	EV: {The absolute ruler Louis XIV ruled France and lived in a similar way [DS: as the picture describes]. [RE: He tried to get the most out of his country by paying low wages which was possible, because of the low prices for corn and cattle. In addition to that, accommodation of money in the country allowed him live the way he wanted.]}
post-task (cycle 1)	

Pre-intervention, EVS03 only described one part of what he saw in a picture without explicitly linking it neither to the concept in question nor to the historical context, leaving it up to the reader to make this connection. Post-intervention, he made explicit how what was shown in the picture matched the historical context, thus presenting a more justified evaluation. Of course, this justification could have been even stronger, first, by giving a clear verdict and, second, by being more precise when including the source's content.

Furthermore, the learners of group A who could justify their answer better often provided enough detail and did not miss the main point, leading to high results concerning *scope of content*. Statistically speaking, *scope* showed an almost perfect correlation with *justification/comprehensibility* ( $r_p = .90$ ,  $p < .001$ ,  $n = 16$ ) and a strong correlation with *systematicity* ( $r_p = .81$ ,  $p < .001$ ,  $n = 16$ ) and *accuracy/relevance* ( $\tau_b = .73$ ,  $p < .001$ ,  $n = 16$ ). Analogous to the increase of *accuracy/relevance*, one could explain the rise in *scope of content* with higher retention rates when engaging with one topic more extensively. Nonetheless, the mean values of *scope* and *justification/comprehensibility* were the only two values below 2.0 at T2, meaning that there was still considerable room for improvement concerning these aspects.

Turning to the two best results of the first sitting, namely *target level* and *target competence*, one can observe the following: *Target level*, measuring the extent of performing on the intended cognitive level, and *target competence*, assessing the extent of performing the intended historical competence, only present small gains. This could potentially indicate that only the (subject-specific) aspects focused on in the intervention were positively affected in the students' productions.

#### b) Linguistic rating

As Figure 25 shows, the students' productions in terms of language improved on all scales. However, only the differences for *linking in terms of function* and *form* are statistically significant and present a medium effect size as measured by Wilcoxon signed-rank tests (*function*:  $T = 58.0$ ,  $p = .022$ ,  $r_w = .43$ ,  $Mdn_{T1} = 1.0$ ,  $Mdn_{T2} = 2.0$ ; *form*:  $T = 74.5$ ,  $p = .032$ ,  $r_w = .38$ ,  $Mdn_{T1} = 1.0$ ,  $Mdn_{T2} = 2.0$ ). When combining all language-related descriptors, however, results are also statistically significant, as already mentioned on p. 222.

From a qualitative perspective, some evidence of improvements can be reported as well. For example, LES02, who did not use a single hedging device in her pre-intervention performance, now qualified three of her evaluations (item 4 and 5) as well as her exploration of potential motives (item 2), e.g., by using modals in present and past tense, such as “might” and “may have drawn”, as well as conditional clauses and clauses like “I don’t think that” instead of absolute claims. Yet, it should be noted that the average for *hedging* was the lowest score in the second data set, meaning that there was still considerable room for improvement.

*Nominalisation*, on the other hand, presents the highest score in the T2 data set. In the post-intervention task, no performance was rated as level 0 in terms of *nominalisation*, and level 1 was only assigned three times, cutting the

previous number of level-1 ratings in half. At the same time, five students received level 3, which are three students more than in the first round. While the differences are not statistically significant ( $T = 51.5$ ,  $p = .058$ ,  $Mdn_{T1} = 2.0$ ,  $Mdn_{T2} = 2.0$ ), a medium effect size ( $r_w = .30$ ) can be reported. What is quite noteworthy, qualitatively, is that most learners used a very nominal style whenever they described the workings of a mercantilist system, reflecting the high density of nominalised phrases in the flowchart used during the intervention. IJT12, for example, did not use nominalised phrases or gerund constructions except when she wrote about mercantilism:

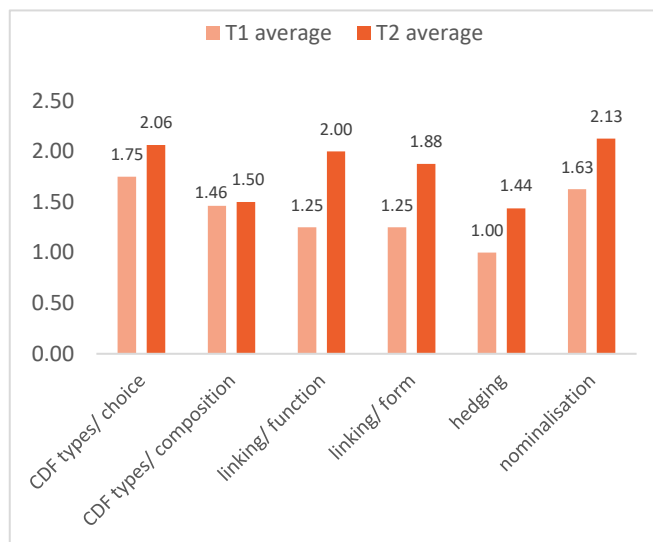


Figure 25. Linguistic rating: T1 vs. T2, group A, cycle 1

89	RE: {The king in France was Louis XIV. Also called the “Sunking” and they had mercantilism in France, [DF: which means that they wanted to achieve an active <b>trade balance</b> , with <b>having</b> more <b>exports</b> than <b>imports</b> .][EA: This was achieved with cheap materials from the colonies, low wages, which were possible, because of the cheap <b>production</b> and also low <b>transport costs</b> which were possible, because of the <b>extension</b> from the roads and channels in the country.]}
Item 3	
IJT12 (A)	EA: {He or she produced this to show the people that the country was exploiting the colonies and that they had to do something to help themselves. [EO: It could also be a part of the (Aufklärung) which showed the people, that they have to begin to thin and not believe in everything.]}
Item 2	

While IJT12 used a range of nominalised phrases whenever she wrote about mercantilism, which she potentially memorized by heart on the basis of the flowchart, she did not use any nominalised phrases in item 2 when writing about something she could not have studied for. This insight implies that it does make sense, especially for weaker students, to provide them with linguistic building blocks they can then use in their own production.

Coming back to cohesion, as mentioned before, results for *linking in terms of form and function* show statistically significant increases, with *linking in terms of function* presenting the biggest



absolute difference between T1 and T2. One reason for this development is that learners linked their ideas more frequently after the intervention. For example, HIP11, who did not use a single linking device in the first assessment, now appropriately used “because of”, “in addition”, “furthermore”, and “although”. Moreover, fewer students used linking devices that appear to be linguistically inaccurate (orthography, syntax, lexical choices) or unfit for their communicative intention. Nonetheless, there were still many instances of learners using inappropriate linking devices after the intervention. For example, OPB04 used “because” to add something she seemed to remember from class, as this information was not visible in the source and did not present a fitting cause-effect relationship:

90	DS: {The picture shows the wealthy mother country who “sits” on the table and waits for the colonies to deliver their products like for example raw materials, foodstuff, gold & silver [RE: <b>because</b> they had to deliver it for a very cheap price].}
Item 1	
OPB04 (A)	

Furthermore, a description per se, as she presumably intended to do, would not require a justification or a reason.

Overall, however, most learners presented a positive development, such as NNM05. This student often used a linking device unsuitable for the intended linguistic function in the first written task (thus receiving 0 points for *linking in terms of function*), and now she employed at least one linking device per item in line with her communicative intention:

91	EV: {I think that this picture is true <b>because</b> it was hard for the colonies to get accepted and nearly all of them had to live somewhere else. [EA/RE: <b>The main reason why</b> mercantilism was established was that the ruler needed money and with the state-running system he gained money. Domestic production increased to have more exports than imports which was the main goal].}
Item 4	
NNM05 (A)	

What is interesting, here, is that NNM05 used appropriate linking devices as well as a CDF composition others were rather successful with for answering item 4, and at the same time, she also presented either completely inaccurate (first sentence) or irrelevant (remaining sentences) information. In other words, cohesion did not necessarily result in good answers. Moreover, some learners even included too many linking devices, overcomplicating the structure and message of their answer, as illustrated by sample 92:

92	DS: {In this picture we can see the colonies (probably of France) giving the state their raw materials and resources to the absolutistic king/queen for free. [EA/ RE: <b>Since</b> and absolutistic leader had the power over a state all for himself and what he said was hole (holy) <b>since</b> he was divine.] <b>Even though</b> the mother state has food on the table it still wants more and more not caring about what the others have.}
Item 1	
EVA02 (A)	

The structure of this answer is a bit nested, with two dangling explanations. In addition, these explanations were not visible in the source and were probably reported based on what the student remembered from class.

Looking at other data types, the students’ increased use of linking is likely to stem from the teacher’s insistence on their use during the lessons, resulting in more awareness concerning the

necessity of linking. For example, sometimes the teacher gave feedback in this respect, e.g., when the students did not use any phrase or when they used too many phrases. The extract below serves as example of the latter case:

**Extract 93, lesson 3, cycle 1:**

- 1 ETS12: Ähm, **due to** this situation his ministers, mistresses and everyone else knew his weakness **resulting in** spoiling and X the king, **since** this was XX.
- 2 TA: So once again, **very long sentence**. Once again and louder.
- 3 ETS12: **Due to** this situation his ministers, ministers, mistresses and everyone else knew his weakness **result**
- 4 TA: = **Punkt [Full stop]**. Okay?

In this extract, the teacher comments on ETS12's very long and nested sentence, which contains three phrases of the language box. When asked to repeat her sentence, T<sub>A</sub> interrupts her, showing her where to split the sentence.

Additionally, the teacher gave feedback when they used phrases without including the required content:

**Extract 94, lesson 3, cycle 1:**

- 1 EVA02: As a result äh, did the ministers have very high privileges and
- 2 TA: **As a result of what?**
- 3 EVA02: Of. äh, the hi-, knowing his weakness like using it.
- 4 TA: = Yeah. Exactly. So, but **you have to include this information somehow**.

Turning to the use of CDF types, *choice of CDF types* moderately improved from the first to the second sitting. This difference is not statistically significant ( $T = 17.0$ ,  $p = .163$ ,  $Mdn_{T1} = 2.0$ ,  $Mdn_{T2} = 2.0$ ), but it does present a small effect size ( $r_w = .25$ ). At T2, only three performances were rated as level 1 for *choice of CDF types* ("some CDF types are target CDF types"), compared to six students at T1. Additionally, four students reached level 3 ("all episode and most basic CDF types are target types") in the second sitting, doubling the respective count of the first sitting. Two of these students (ELF03 and OPB04) even jumped from level 1 to level 3. In their first performance, these students mainly reported facts they had memorized without linking them to the source; thus they did not EVALUATE or EXPLORE as was asked of them. After the intervention, they engaged more directly with the source, paying more attention to the prompt:

95	EV: {In my opinion the source depicts it truthfully because [RE: the colonies had do deliver cheap raw materials to the mother countries][DS: as it is shown in the picture. What is more is that the mother country looks really rich in the picture and the colonies look very poor][EV: which was also the truth. But of course it is only one part of mercantilism.]}
Item 4	
OPB04 (A)	

Here, OPB04 did not only report contextual information, but she also evaluated the validity of the picture by relating reported facts to what was shown in the picture. Similar trends could be observed in other students, too, moving away from mere reporting to explicit evaluations.

When it comes to *composition of CDF types*, most of the issues observed in the needs analysis remained unchanged. It should be noted, however, that three learners did not use more than one CDF type per item in the first assessment, and thus *composition* was not rated then. OPB04 above

is such an example, jumping from no rating in terms of *composition* to level 2. Those that were rated in the first round mostly remained on their level (six students) or even descended one level (four students). Only three students improved from the first to the second sitting. One issue present in the performances both at T1 and T2 is that some students' productions contain a range of different CDF types that do not seem to build on each other, thus leading to an unclear superordinate communicative intention, i.e., CDF episode. Sample 96 serves as an illustration:

96	[EA: To display the mother countries power.] [DS: "She" is not doing anything good for them still they are chained like slaves and have to give up on their goods. [CA: Looking at their faces they look happy DS] [RE: but actually it is just harming them]] [CA: [EO: because they could develop][RE: but instead their progress is very slow]]. [RE: (No money to invest into development and no goods to sell to receive money)].
Item 2	
ARJ01 (A)	

This answer is very nested, and it is unclear what ARJ01's overall intention was. The student started off with an explanation of the artist's motives, as specified in the task, and then he seemed to contrast what could be seen in the picture (DESCRIBE) and what the colonies experienced (REPORT) or could have done if they had not been exploited (EXPLORE). The last REPORT sequence could also be meant as an explanation why the colonies' progress was slow, but since he did not mark it linguistically as such, there is no way of knowing whether this was his intention. Overall, it remains unclear what his main communicative intention was.

### c) Summary and implications for the design

The students' post-intervention productions of cycle 1 indicate that the learners improved both in terms of content and language in the areas targeted by the intervention and measured by rubrics designed for this study. Consequently, it can be argued that the learners' written production was indeed positively affected by the intervention to a considerable and often statistically significant degree.

Zooming into different categories of the rubrics, one can see that those areas with particularly low ratings in the pre-intervention task could show considerable growths, such as *accuracy/relevance, justification/comprehensibility, systematicity* and *scope* on the content scale and *linking in terms of form and function, hedging*, and *nominalisation* on the language scale. In these areas, substantial headway had been made, although some issues still remained, yet to a lesser extent. Concerning nominalisation and systematicity, similar findings were reported by Lo and Yeong (2018) while Breeze and Gerns (2019) observed improvements in terms of linking in their intervention study.

At the same time, only little change could be noticed when it comes to *target level, target competence*, and *composition of CDF types*. While *target level* and *target competence* already started out on a high level, *composition of CDF types* remained on a relatively low level. It appears that creating a logical montage of CDF types that support or build up to the main target CDF type (i.e., CDF episode) still posed a problem for many learners of this group. However, this is not surprising, as even more advanced learners struggle with well devised and linked CDF

compositions, as Breeze and Dafouz (2017) point out. As a consequence, it is all the more important to focus on these aspects in upcoming cycles.

Finally, looking at areas receiving high ratings in the first round (i.e., *target level* and *target competence*), one lesson of the first cycle could be not to overlook aspects that initially seemed solid. Thus, these areas should be considered more in the upcoming cycles, e.g., by new language boxes, more concrete language boxes, or in the form of awareness-raising remarks in the teacher's version of the materials. This way, teachers might notice these issues and potentially discuss them with the learners.

#### **7.4.2 Cycle 2: absolutism and mercantilism (school B)**

The unit on absolutism and mercantilism was revised based on the insights gained in cycle 1 and adapted to fit the context of school B. Four days after finishing the implementation of the revised materials, the students of group B had to complete the post-intervention written task, and a smaller group of students shared their views in a retrospective interview. The students who were absent due to a language contest completed the written task a day later in another teacher's lesson. Here, we also attached a short written open-ended feedback form to the task (see [appendix section I/ E](#) and subsection 5.3.3.4) in order to also hear the voices of students with high linguistic aptitude (EBF05, UCQ07, UKV05, ABS04). These responses are also considered in the following subsection. As for the composition of the focus group, EBF05 (who participated in the language contest) was replaced by HRG10, who volunteered. The interview with the teacher took place two months later due to his tight schedule but also to allow us to discuss (preliminary) outcomes of the student interview and the post-intervention task, contextualizing these results and gauging their plausibility from a teacher's perspective. Additionally, in this section, relevant insights gained through lesson observation or examining student worksheets<sup>94</sup> are connected to the students' interview, written feedback, or written task performance whenever appropriate.

##### **7.4.2.1 Retrospective interviews with students and teacher**

###### **a) Experiences with the intervention**

For the teacher, it was mostly a positive albeit quite different experience to his normal lessons in the sense that he had never used English to this extent and had never implemented CLIL to this degree. Thus, he had to prepare quite differently, including preparing vocabulary and reading on the topic in English, which, he added, he enjoyed. Although he was used to learner-centred teaching and scaffolding his tasks, he had never been as explicit about communicative intentions and precise expression, which is something he really seemed to value:

---

<sup>94</sup> Group B completed their worksheet on Louis XIV on their laptops and handed them in digitally. These documents were anonymized (using the learners' codes) and made available to the researcher by T<sub>B2</sub>.

97	English translation	Original quote
T <sub>B2</sub>	your materials were really, <b>really precise</b> and focused partly on trying to <b>express things precisely</b> . And this is something I normally didn't implement as much, so this is something that <b>I've taken away from these materials, more than anything else</b> .	<i>deine Materialien waren sehr, sehr genau und fokussiert teilweise auf ein, ähm, dass sie versuchen, Dinge wirklich auch genau zu formulieren. Und das ist ja etwas, was ich sonst sicher nicht so stark eingesetzt habe, also das ist auch etwas, was ich mir aus diesen Materialien am meisten mitgenommen habe.</i>

Later he stated that being more explicit concerning communicative intentions and precise expression, including paying attention to how one could mark linguistically whether something was fact, opinion, or assumption, was something he would definitely consider more in his teaching, be it in English or German. Unlike group A, the students of group B did not really talk about any differences in linguistic practices by the teacher. In the implementation phase, T<sub>B2</sub> only rarely insisted on the students' use of phrases and only sometimes commented on the students' language use. Here is one example:

**Extract 98, lesson 4, cycle:**

- 1 TB2: Perfect. Äh, so, who will tell us his or her definition. [...]
- 2 USN05: Mercantilism is, is äh, not relying on foreign import but relying only domestic production and
- 3 TB2: = Okay, this is very good, it's right of course, **it's not wrong. But is it a good definition?**
- 4 USN05: No, it's
- 5 TB2: What would be probably a better one? Could be a better one?
- 6 AMM01: Ähm, mercantilism is a state-run economy with the aim of an active trade balance with more exports than imports.
- 7 TB2: = Yeah, perfect. So, components of a definition. **You have broader category and then a more specific description. So, for example** mercantilism can be defined as a state-run economy aimed at an active trade balance, which means that there should be more exports than imports. So, quite easy if you try to write down a definition start with the broader category what is it about and then a more specific description. Some further information examples and so on. **It's quite the same problem with my younger boy, äh, he just, äh, always tells me something and he always start with a very specific information, and I have no idea what he is actually talking about.** So always start with the broader category and then we go into further details.

While accepting the content of USN05's answer as correct, T<sub>B2</sub> questioned whether this could be considered a definition (line 3), asking for a better example, which AMM01 was able to deliver (turn 6). The teacher praised AMM01's input and continued with explicit information on the constituents of a definition, then repeated a version of AMM01's answer, and finally related it to everyday language use by telling an anecdote. The students of group B did not talk about these sequences in their interview, neither seeming to oppose nor to actively appreciate these episodes, which indicates that the learners accepted them as part of T<sub>B2</sub>'s teaching. Overall, there were not too many of these episodes in the five lessons observed, but there were quite many instances of clarifying terminology, often using German in the process. The only aspect the student interviewees noticed concerning linguistic practices was the increased use of English, which was quite a change for them. Some learners wished for more episodes in German while the language-contest-students filling in the feedback form appreciated the rather consistent use of English.

The biggest differences besides the more prominent use of English were the workload and intensity of the history lessons:

99	English translation	Original quote
ICK01, OVD11, EOS12 & IMJ07	R: To start with, I'd like to ask you which differences you noticed compared to your usual lessons? OVD11: We did a lot <b>more than usual</b> . ICK01: = Yes = IMJ07: Rather more done <b>in English</b> . OCD11: Yes. EOS12: We <b>did a lot more in general</b> (laughs). IMJ07: More <b>exhausting</b> . Sfx: <b>A lot</b> .	<i>R: Ganz am Anfang würd ich euch gern fragen, welche Unterschiede euch denn im Vergleich zu regulären Stunden aufgefallen sind?</i> <i>OVD11: Wir haben viel mehr gemacht als normal.</i> <i>ICK01: = Ja =</i> <i>IMJ07: Eher mehr auf Englisch gemacht.</i> <i>OVD11: Ja.</i> <i>EOS12: Wir haben <u>überhaupt</u> mehr gemacht (lacht).</i> <i>IMJ07: Anstrengender.</i> <i>Sfx: Viel.</i>

This is something T<sub>B2</sub> observed too, stating that he would normally not plan such dense, labour-intensive, and detailed units. He reported that he tried to include frequent brief recaps in German to ease the cognitive load and to ensure his students understood. He further tried to solve the problem of exhaustion paired with loss of interest by connecting the historical content to current politics by means of whole-class discussion. Still, the learners were somewhat overwhelmed by this unit. However, they also seemed to appreciate CLIL now more than they used to, considering their quite negative attitudes in the needs analysis interview. Yet, while they would like to experience more CLIL, they would prefer it in a less intense manner than in this intervention:

100	English translation	Original quote
IMJ07, ICK01 & OVD11	R: Potentially as a closing question, would you like to have more CLIL every now and then? IMJ07: Yes. R: On principle? ICK01: Not as intense as the last two three weeks, but [...] yes. OVD11: Increasingly, yes.	<i>R: Vielleicht als Abschluss, hättet ihr gerne mehr CLIL manchmal?</i> <i>IMJ07: Ja.</i> <i>R: Prinzipiell?</i> <i>ICK01: Nicht so intensiv wie jetzt die letzten zwei drei Wochen, aber [...] schon.</i> <i>OVD11: vermehrt schon.</i>

Interestingly, the four students who participated in the language contest and thus gave their feedback in written form did not mention the workload at all. These linguistically gifted students appeared to be undeterred by labour-intensive materials. Instead, their comments indicate a mostly positive experience, welcoming the consistent use of English and the opportunity to work autonomously.

## b) Evaluation of the intervention

Bearing in mind that teacher and students experienced this unit quite differently, their evaluation of the intervention varies too. The code co-occurrence models (Figure 26 and Figure 27) indicate which aspects were mentioned how often and whether these points can be regarded as positive, negative, or inconclusive evaluations, considering both the interviewees and the four students filling in the short questionnaire. While the teacher spent much more time talking about aspects he appreciated (17 positive, 6 negative, 5 inconclusive codings), the students' evaluation seemed more mixed, with 31 positive, 27 negative, and 8 inconclusive codings. Here it should be noted that the written feedback as well as the interviewee's red and green markings directly on the materials, which indicate negative and positive evaluations, paint a much more positive picture than the actual focus interview.

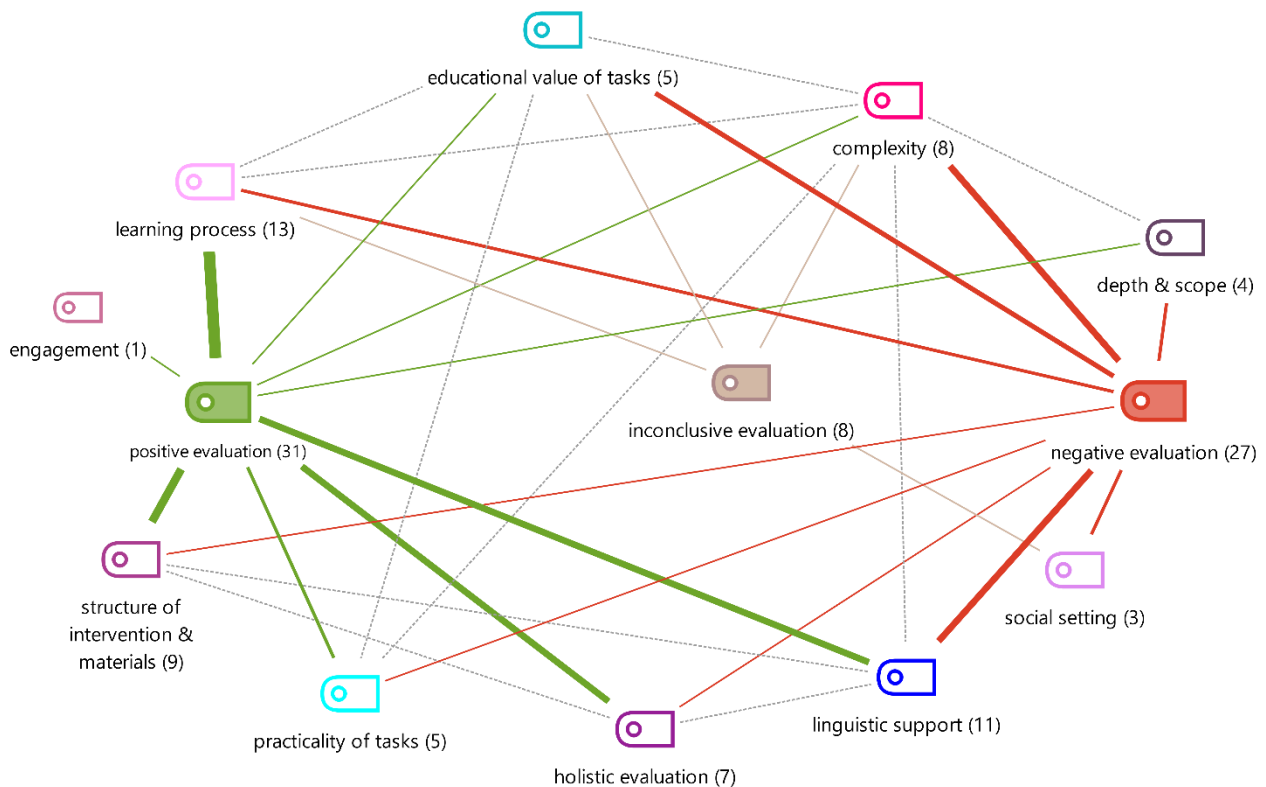


Figure 26. Code co-occurrence model: evaluation by the students of group B

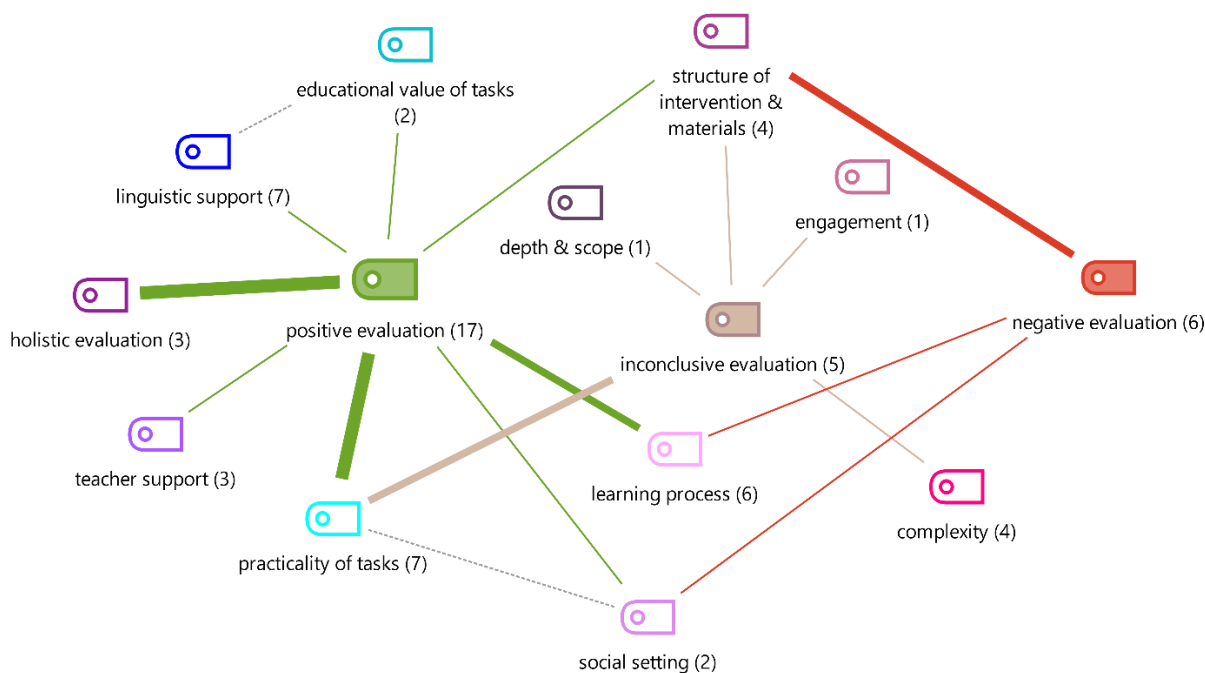


Figure 27. Code co-occurrence model: evaluation by T<sub>B2</sub>

Beginning with **engagement**, this code was only given one time each, suggesting that this was not an aspect that stood out for the participants to be different to their normal lessons. When asked explicitly, the students said that they were more active than usual. The teacher, however, considered the situation in a more nuanced way. From his point of view, it engaged those learners more that were already more proficient in English, as they would sooner experience a sense of achievement. Less proficient learners, he assumed, could be discouraged, thus participating less. Looking at the worksheets on Louis XIV, which the learners handed in after the unit, this estimation can be tentatively confirmed. For example, students reporting rather low grades the previous year (e.g., ARC11, EOS12, HRG10, or USN05) handed in incomplete worksheets, suggesting that they did not engage with the materials to a great extent. In USN05's case, interestingly, active oral in-class participation could be observed whenever he was present (four out five lessons).

Turning to **complexity**, interviewees mostly rated this aspect as negative, as they often felt overwhelmed due to long and difficult texts and input, including the flowchart, which seems quite different to the experiences reported by group A. In the written feedback, however, nobody reported any issues concerning the level of difficulty. This mostly matches T<sub>B2</sub>'s observations that the materials stretched the limits of low achievers while not boring high achievers. Overall, he rated the material's complexity, regardless of the language, as moderately high. In relation to high levels of complexity and density of content, the interviewees expressed feelings of exhaustion. T<sub>B2</sub> also took notice of this, yet he argued that one could expect students to put up with demanding lessons from time to time, as explained in the following extract:



101	English translation	Original quote
T <sub>B2</sub>	Well, it was taxing, yes, on the other hand, <b>it should be allowed to be taxing</b> . Maybe that's something that's still in our heads 'lesson always have to be interesting, exciting and fun', but lessons can also just be exhausting sometimes. <b>And I think one can expect this from them.</b>	<i>Naja, es war schon anstrengend, ja, andererseits es darf ja auch einmal anstrengend sein. Also das ist auch etwas, was vielleicht in Köpfen noch, noch drinnen ist, Unterricht muss immer interessant, spannend, lustig sein, aber Unterricht kann auch einmal einfach anstrengend sein. Und ich glaub, das ist ihnen auch zuzumuten.</i>

Tying in with increased levels of participation and complexity, it was asserted both by learners and teacher that students indeed learned something. Yet, views differed as to which areas improved (most). While some learners felt that they only learned something about history, others claimed that they could take something away both in terms of content and language. The teacher elaborated on this, arguing that the greatest added value in terms of **learning outcomes** was of linguistic nature. When asked about if or how the intervention affected their English skills, some learners confirmed a positive impact but argued that these gains were not necessarily useful:

102	English translation	Original quote
IMJ07 & ICK01	IMJ07: I'd say no because they are not words you use in everyday life. [...] Well, I do think that our vocabulary expanded through all these historical facts that we had the chance to acquire. But for everyday use, I'd say same level. [...] ICK01: I agree.	<i>IMJ07: Ich würd sagen nicht, weil es sind ja nicht Wörter, die du im Alltag benutzt. [...] Also ich glaub unser Wortschatz hat sich erweitert durch diese geschichtlichen Fachkenntnisse, die wir jetzt auf Englisch erworben durften, erwerben durften. Aber sonst so im Alltag, glaub ich selber Stand. [...] ICK01: Ich schließ mich der Meinung an.</i>

Here group A and B seem to be in opposite positions. Group A only appreciated subject-specific linguistic input while rejecting general language tips in the context of history lessons, whereas for group B the situation seems somewhat reversed. The reason for this discrepancy might lie in the difference of their programmes. While group A experienced BE across many subjects and hence a great amount of exposure to the language, group B had not received much CLIL instruction prior to the project and also seemed somewhat dissatisfied with their English lessons in the needs analysis interview. As a result, they might prioritize 'general' English language skills over subject-specific ones.

Apart from learning outcomes, T<sub>B2</sub> offered more insights into how the materials affected the **learning process**:

103	English translation	Original quote
T <sub>B2</sub>	Well, I believe that the type of tasks was very, very focused on one learner type, the <b>visual learner type</b> [...] because you mostly used texts and images. There was <b>not a lot of communication</b> [...] It was intended that they would again and again talk about their results and that also happened some of the time, but sometimes, I think, it was rather short and for them already completed [beforehand].	<i>Also ich glaube es war, ähm, von den Aufgabenstellungen her doch ganz, ganz stark auf einen Lerntyp, auf einen visuellen Lerntyp, der, äh, fokussiert [...] weil sehr viel über, über Texte, Bilder und so weiter gegangen bist. Ähm, es waren jetzt nicht wahnsinnig viel Kommunikation [...]. Es ist vorgesehen, dass sie dann immer wieder drüber gesprochen haben, ähm, es ist auch teilweise hat's stattgefunden, manchmal glaube ich war's relativ kurz und, und, und für sie [bereits] abgeschlossen.</i>

Here, T<sub>B2</sub> explained that the materials mostly targeted visual learners, while communication was somewhat neglected. Even though frequent **interactive activities** were planned, this was sufficiently done only by some learners some of the time. Since they could theoretically work on all tasks on their own, including those tasks that prompted them to pair up, they did not feel the need to really engage with one another. Comparing answers was again superficial despite the teacher's instruction to write down the other person's name for each task, as the following excerpt illustrates:

**Extract 104, lesson 2, conversation between LED08 and EOD03:**

- 1 LED08: *Sollen wir diskutieren? [Should we discuss?]*
- 2 EOD03: *The answers X. ((LED08 reading on EOD03's laptop, and EOD03 reading on LED08's laptop))*
- 3 LED08: **Okay, das heißt [Okay, this means]**
- 4 EOD03: *X, 2 to 3 hours, spend his private time to work on matters of the state, X with different persons, could work to X at any time.*
- 5 [...]
- 6 LED08: *So, he believes that he set a certain time of, between, of two times three hours working time a day, or which do not*
- 7 TB2 [to whole class]: = So remember if you talked about task A, äh, just go on with task B. And after you finished task B, you have to talk to somebody else. So not the same person twice, please change your partner. Switch partner, change partner.
- 8 EOD03: **I think he didn't believe it.**
- 9 LED08 [reading from her laptop]: *He believes that he settled a certain time of 2 times 3 hours working time a day for, which do not include his private effort like overthinking and always being available for urgent matters. Once in a while he spoke publicly.*
- 10 EOD03: **Right X.** ((LED08 and EOD03 reading silently on their own laptops))

The extract above exemplifies how the learners compared their answers in pairs. Mostly, LED08 and EOD03 were reading out their answers, sometimes rephrasing their output (e.g., line 3) or briefly commenting on each other's suggestions (e.g., line 10). However, there was no genuine discussion or sharing of views apart from EOD03's thought in line 8, which LED08 did not acknowledge. Instead, she just read out her answer. This issue also came up in the student interview:

105	English translation	Original quote
OVD11	But <b>comparing was also unnecessary</b> because everybody just says, 'yes, what do you have here', 'yes, okay, roughly the same', so that wasn't really purposeful.	<i>Aber das Vergleichen war auch unnötig, weil jeder so sagt, ja was hast du da, ja okay ich hab eh zirka das gleiche, also das hat sich nicht so viel gebracht.</i>

In the written feedback, the students liked that they could work on the tasks individually before pairing up, but one student mentioned that "the talking was unnecessary". At the same time, in this group, they did not complain about varying work paces like group A did, although the learners of group B did differ in this regard according to the teacher. Thus, for the teacher, the individualisation strategy to only make three out of four sources obligatory and having them first work individually and then pair up with somebody of the same work pace still seemed reasonable but would need more practice and more readiness on the students' part. He further added that the communicative part would need to be more purposeful and indispensable, e.g., by setting up tasks

as information gap activities. So, in terms of **overall structure**, T<sub>B2</sub> would insert a more communicative and creative task in the middle of the materials to address other types of learners, thus increasing variety and, hopefully, general levels of motivation. One student who filled in the survey highlighted the variety of tasks in this unit, but two others found the overall structure boring. As for the structure and sequence of individual tasks, the learners appreciated the small steps and the clear sequence of the tasks, especially of the flowchart worksheet. Only HRG10 found the structure illogical but could not explain why. His peers disagreed:

106	English translation	Original quote
HRG10, EOS12, OVD11, & IMJ07	R: What should one do differently? [...] HRG10: The tasks should be <b>better structured</b> . EOS12: XX, well, it <b>was well structured</b> . [...] OVD11: <b>By and large, I think so too</b> . IMJ07: <b>I also found it good</b> .	<i>R: Was sollte man anders machen? [...]</i> <i>HRG10: Die Aufgabenstellungen strukturierter machen.</i> <i>EOS12: XX, also war <u>gut</u> strukturiert. [...]</i> <i>OVD11: Im Großen und Ganzen fand ich das eigentlich auch.</i> <i>IMJ07: Ich fand's auch gut.</i>

As for the **practicality** of tasks, most of the learners were happy about working with digital files, but some of the additional symbols, like the magnifying glasses for language tips turned out to be an obstacle when filling in the digital files. Concerning clarity of tasks or prompts, most students were satisfied, highlighting the colour-coding of the mercantilism worksheet. The teacher agreed with the students' evaluation. He added that the most frequently asked question during the implementation did not concern the tasks per se but the amount or the nature of the written output required and sometimes also the sequence of the tasks. However, he thought that this could hardly be avoided unless you gave them very detailed prompts, but this might deter them from the start. Turning to the practicality of the materials from a teacher's point of view, T<sub>B2</sub>'s remarks concerning teacher support were entirely positive, describing them as clear and helpful.

Another aspect the teacher assessed completely positively was the **linguistic support** provided for the students. As mentioned already, this is something he said that he would implement more in the future. He also appreciated that the boxes often offered concrete examples, illustrating the meaning of the language tips. When informing the teacher that the more proficient students of group A felt patronized by the linguistic support, he argued that, as a teacher, one should communicate the purpose of scaffolding more clearly. He further explained that once learners felt not to be needing this kind of support anymore, they could just happily ignore it, as the ability of doing it independently would form the endpoint of scaffolding after all. Nonetheless, the teacher should still monitor whether these learners had actually already achieved that end point:

107	English translation	Original quote
T <sub>B2</sub>	TB2: <b>Honestly, I'd be definitely okay with that, because, in the end, this is scaffolding.</b> The moment when you feel, you don't need the scaffold, <u>okay</u> . R: Then you don't use it.	<i>TB2: Ich kann ehrlich gesagt ganz gut damit leben, weil ich mein im Endeffekt das ist ja Scaffolding. In dem Moment, wo du das Gefühl hast, du brauchst das Gerüst nicht, <u>okay</u>.</i> <i>R: Dann nimmt man's nicht.</i>

	TB2: Then you don't use it. If the output might still not be as strong because the language is not precise enough, one could still point towards this, 'well, but there are, you still have some weaknesses'.	<i>TB2: Dann nimmst du's nicht. Ähm, wenn der Output dann aber vielleicht nicht ganz so stark ist, weil eben die Sprache nicht genau genug ist, dann kann man ja durchaus noch einmal drauf hinweisen, naja aber da gibt's, du hast noch Schwächen</i>
--	---	--

The learners' opinions about linguistic scaffolding, in contrast, were more mixed. While some argued that the boxes helped when working with challenging input, others did not feel that the linguistic support was useful to them, as illustrated by the following examples:

108	English translation	Original quote
ABS04 [WF] <sup>95</sup>	Language boxes <b>unnecessary</b> (in my opinion)	<i>Language boxes unnötig (in my opinion)</i>
ICK01	It got more and more complicated and at one point, <b>the phrases in the box didn't help a lot anymore, and from that moment onwards, I didn't get it any longer.</b>	<i>Es wurde dann halt immer komplizierter und irgendwann haben die Phrasen dann in der Box auch nicht mehr viel geholfen und ab da hab ich's dann nicht mehr verstanden.</i>
OVD11	I did read through it [= language boxes], but <b>it wasn't really useful.</b>	<i>Ich hab's [= language boxes] mir schon durchgelesen, aber es hat sich nicht viel gebracht.</i>
UKV05 [WF]	<i>[What did you like in particular about the CLIL lessons on absolutism and mercantilism?]</i> <b>UKV05: There were always tips and words one could use</b>	<i>[Was hat dir in den CLIL Stunden dieses Projekts (Absolutismus &amp; Merkantilismus) gut gefallen?]</i> <i>es gab immer wieder Tipps welche Wörter man benutzen kann</i>

For some learners, like ICK01 quoted above, the phrases were actually crucial for understanding the content. However, at some point, the content was perceived as too complex despite the help provided by the linguistic support measures. The same student later clarified that boxes that were concise and quite specific to the task were especially useful to avoid overload. Amongst the interviewees, there was no consensus as to whether phrase boxes or explicit language tips were more useful.

Looking at their worksheet productions, almost everyone used the phrases provided, especially when structuring longer texts. Nominalisations were also widely used, yet some nominalised phrases were, of course, copied and pasted from the text. Here is an example:

109	Duke criticised that Louis' <i>jealousy</i> became in a certain point a <i>weakness</i> for him. He also says that Louis didn't choose the ministers <u>because of</u> their <i>knowledge</i> or <i>capacity</i> , but <u>because of</u> their <i>ignorance</i> and their <i>want of capacity</i> . <u>As a result</u> of his <i>vanity</i> and <i>wanting</i> to be admired, he ruined himself.
task B/1 worksheet	
NGS01 (B)	

NGS01's writing style can be described as quite nominal. It should be noted that all of these words can also be found in the input text the students had to work with (marked in *italics*). Still, the nominalised words are mostly integrated well into her own sentences, which feature some of the phrases provided in the language box alongside this task (underlined), too. As a first step, this

<sup>95</sup> [WF] indicates that this quote was taken from the students' written feedback.

might be a good way of helping students incorporate features of historical discourse. Something similar could also be observed during their final reflection in lesson five, when T<sub>B2</sub> used the app *Padlet* to revise the last four lessons, using two guiding questions. Again, many learners typed in nominalised phrases they picked up from the worksheets, such as “abolishment of domestic tariffs”, “centralized administration”, or “more exports than imports”. However, they also came up with their own terms, such as “injustice”, “dictatorship”, “elegance”, “authority”, or “exploitation”. In the worksheet productions, hedging could be observed less often than nominalisation. Only answers to certain tasks showed these features more regularly, namely especially those tasks that were accompanied by a language box on this topic. The extract below serves as an example of a student whose pre-intervention task performance was well below average, both in terms of content and language, but who nonetheless managed to incorporate linking devices, nominalisations, and hedging into her in-class writing:

110	Historian <u>seem to agree</u> in some aspects that Louis XIV was a great leader and king, who built foundations and other things like this. <u>But we know from other resources</u> that he was the chief architect of <u>royal tyranny</u> . <u>so from my point of view</u> its very clear that Louis was a great leader and King who had enough power and people by his side who supported him.
ZE11 (B) final task worksheet	

ZE11 used many phrases provided by the material (underlined), but from a content perspective, her assessment appears somewhat fragmentary. To be more precise, her final conclusion does not reflect the contrast she elaborated on for most of the paragraph. Therefore, the concluding “so” does not seem quite appropriate in this context. In general, weaker students had problems to organically integrate the tips and phrases into their own writing, suggesting that this process would need more time. When appraising the intervention holistically, T<sub>B2</sub> expressed a similar thought:

111	English translation	Original quote
T <sub>B2</sub>	I believe that it is a bit of a <b>learning process</b> that language is simply central [...] I believe that it [ = the intervention] is a <b>beautiful building block</b> towards that goal, but it's not something where you could say [...] ‘yeah I worked through this one and now [I’m done with] this topic until A-levels. If you’ve got <b>more of these building blocks</b> and use them <b>regularly</b> , then I definitely believe that <b>they</b> [the students] <b>will start using language with more sophistication</b> .	<i>Ich glaub, dass es auch ein bisschen ein Lernprozess ist, dass Sprache einfach zentral ist [...] ich glaub es ist ein schöner Baustein auf dem Weg dorthin, aber nichts, wo du jetzt sagst [...], juhu, hab ich jetzt abgehakelt und bis zur Matura hab ich das Thema [...] wenn man mehrere dieser Bausteine hat und die auch regelmäßig wieder einsetzt, dann glaube ich durchaus, dass sie anfangen, Sprache differenzierter zu verwenden.</i>

Nonetheless, it could be observed during the lessons that learners were able to adequately express their thoughts in English, using phrases provided as well as nominalisations and hedges to some extent, as the following examples illustrate:

#### Extract 112, 113, and 114, lesson 4, cycle 2:

- 112 INY06: By offering governmental loans as well as founding royal manufactories, the state can work towards an increase of production.
- 113 ABS04: Because of protective tariffs, the import would cost way more; therefore, they will think about importing something twice, which leads to the prevention of imported goods. So, green [part of the flowchart] will increase exports, purple [part of the flowchart] decreases imports.

- 114 NTE12: Yes. Äh, active trade balance makes accumulation of money in the country possible, which allowed to enlarge the army, to centralize administration, and have an expensive life at court.

Here, INY06, ABS05, and NTE12 used a very nominal style and phrases offered by the worksheet on mercantilism. ABS04 even tried to express some degree of hypotheticality by using *would* and *will* constructions.

c) Summary and implications for the design

Similar to cycle 1, the overall tone of the interview with students and the teacher differed quite considerably, with the teacher being rather enthusiastic about the intervention, while students seemed more sceptical, which mostly seemed to stem from the high level of labour-intensity. Looking at the learners' red and green markings directly on the materials, it seems that they, in theory, liked most of the tasks, as green markings prevail. Also, the four learners who gave their feedback in written form seemed rather satisfied with the materials. Aspects appreciated both by the student interviewees and respondents of the survey include the type of tasks, clear prompts and colour-coding, the structure of the worksheets (small steps and clear sequence), learning outcomes, and partly also the linguistic support. Holistically, the interviewees put forward that this intervention had increased their interest in CLIL on the condition of decreasing labour-intensity and complexity of input. Overall, the interview as well as the filled-in worksheets suggests that this unit was overwhelming for the mid and low achievers of group B, who had hardly had any CLIL experience prior to the intervention. The linguistically apt students filling in the survey did not report any such issues and also their submitted worksheets do not indicate overload. This is in line with the teacher's observation and ties in with recommendations by Meyerhöffer and Dreesmann (2019), Otwinowska and Foryś (2017), or Somers and Llinares (2018), who all call for adequate and sufficient linguistic scaffolding in order to avoid frustration and demotivation, ensuring that learners would "not only survive through but derive pleasure from productive learning in the CLIL classroom" (Otwinowska & Foryś, 2017, p. 475). In the case of group B, scaffolding for low to mid achievers could have been even more prominent and comprehensive. Some learners relied on this support and felt that, at times, the complexity of the task exceeded the support provided. The learners also suggested including the L1 more often to ease cognitive load, which corresponds to the views of the learners in an intervention study by Meyerhöffer and Dreesmann (2019).

More generally, the learners of group B appreciated the idea of having linguistic support, by and large, and also seemed to consider the language tips and phrases when using language in the classroom (written and oral); yet some students struggled with organically integrating the phrases provided into their own utterances and written sentences. Concerning type of linguistic support, the students of group B favoured linguistic input that was concise and targeted to the task, similar to the pilot groups (Bauer-Marschallinger, 2019). Unlike group A, subject-specificity was not of any concern for the learners of group B. In fact, group B preferred general linguistic support that would also help them in their everyday English use, and some learners even felt that

history-specific support was pointless. Thus, it appears that learners with little CLIL experience and, potentially, low satisfaction levels with their regular English classes appreciate general academic English support, while more experienced bilingual students (like group A) prefer subject-specific input over general language advice. The teacher, in any case, welcomed the inclusion of the language boxes and found those tips most useful that highlighted the purpose of certain performative verbs and how one could express these communicative intentions.

The teacher recommended to increase variety of task types and to appeal to different learner types, including social learners. Although the individualisation approach of having them work at their own pace and then comparing with someone of similar speed did solve the issue of diverging work paces, it did not allow for purposeful in-depth collaboration. So, T<sub>B2</sub> suggested designing information gap activities requiring goal-driven communication. By doing so, one could also cater for different learning preferences and achievement levels. Including more communicative as well as creative tasks would also help reduce the density of the input and the risk of cognitive overload. For CLIL beginners or less proficient English users, it might be more valuable to include elements of this approach regularly but not exclusively and with a diverse set of tasks.

#### 7.4.2.2 Post-intervention written tasks

In cycle 2, the students of group B had to complete the same competency-based written task as the students of group A in cycle 1, which was a caricature dealing with mercantilism (and absolutism to some extent). The differences between pre-intervention written task (T1) and post-intervention task (T2) performance observed in group B ( $n = 19$ ) are visualized in Figure 28. A significant upward trend can

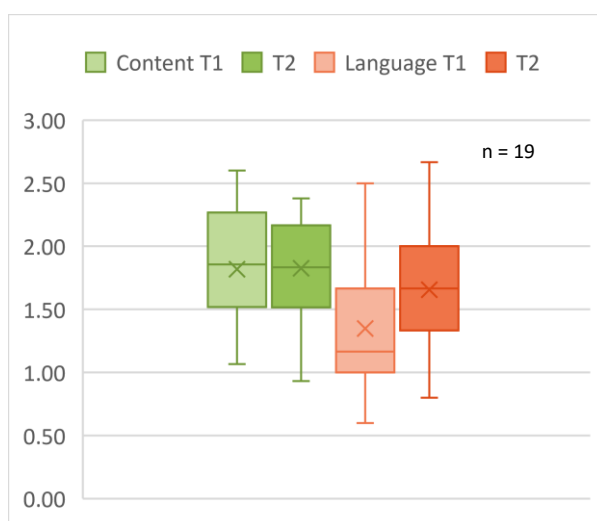


Figure 28. Comparison T1 vs. T2: boxplots, group B

only be observed in terms of language, from  $M_{T1} = 1.35$  to  $M_{T2} = 1.66$ , with  $t(18) = 2.450$  and  $p = .025$  and a large effect size of  $d = 1.00$ . Content results do not show any significant changes ( $M: +0.01$ ,  $Mdn: -0.03$ ), unlike in group A, where ratings in both areas significantly increased after the first round of intervention. In group B, standard deviation decreased from T1 to T2 in both domains (language: from  $SD_{T1} = 0.59$  to  $SD_{T2} = 0.52$ ; content: from  $SD_{T1} = 0.47$  to  $SD_{T2} = 0.37$ ), meaning that the data is less dispersed after the intervention. Interestingly, content results at T2 show a lower minimum and maximum value than at T1, while for language, minimum and maximum values are higher compared to their T1 counterparts.

Looking at individual trajectories (Table 12), ten students achieved lower content results after the intervention, while nine performed better or at exactly the same level. In any case, most of the changes are small, with only one student changing more than a full point (USN05). When it comes to language, however, there are only five learners who performed worse after the intervention, meaning that the language rating of 14 students increased from T1 to T2. Some individual learners improved their language results even by a large margin, such as ZEA11 (+1.83), ICK01 (+1.00), or OVD11 (+0.83), i.e., three students who scored rather low in the first sitting.

Splitting the group into three achievement levels based on their T1 overall performance indicates that students of different proficiency levels (as measured in the first test) developed quite differently. As Table 13 shows, students with low scores on the first sitting improved both their content and language ratings considerably, whereas mid achievers only made some linguistic progress and, on average, received lower content scores. In fact, the mid-achieving group was outperformed by their low-achieving peers in terms of content. High achievers, on the other hand, were not outperformed by any other group, although they lost points on both scales. It is not surprising that students with low T1 ratings were able to gain more than their peers simply because their potential for growth is theoretically the biggest, considering that the scale already

ends at 3.0. Yet, there is no tangible explanation why mid and high achievers actually lost points. If the content had been too complex in English, then the results of low achievers would have

Table 12. Individual results, group B, cycle 2

student code	content			language		
	T1	T2	T2-T1	T1	T2	T2-T1
ABS04	2.50	1.83	-0.67	2.33	2.50	0.17
AKM12	1.96	2.03	0.07	1.67	1.83	0.17
APK08	2.20	1.50	-0.70	1.33	1.00	-0.33
ARC11	1.07	1.47	0.40	1.00	1.33	0.33
ARM03	2.27	1.87	-0.40	2.17	2.33	0.17
DRI04	1.19	0.93	-0.25	0.60	1.17	0.57
EBF05	1.93	1.70	-0.23	1.20	1.67	0.47
EOD03	2.33	2.23	-0.10	2.17	1.67	-0.50
HRG10	1.86	1.41	-0.44	1.00	1.50	0.50
ICK01	1.24	1.97	0.72	0.67	1.67	1.00
IMJ07	1.52	1.73	0.21	0.67	0.80	0.13
LED08	1.61	2.29	0.68	1.67	2.17	0.50
NGS01	2.28	1.52	-0.76	1.17	1.00	-0.17
OVD11	1.52	1.79	0.27	1.00	1.83	0.83
UCQ07	2.60	2.17	-0.43	2.50	2.00	-0.50
UKV05	1.69	1.64	-0.05	1.17	1.33	0.17
USN05	1.11	2.22	1.12	1.00	1.33	0.33
WAS01	2.00	2.00	0.00	1.50	1.67	0.17
ZEA11	1.68	2.38	0.70	0.83	2.67	1.83

Table 13. Differences T2-T1, divided into achievement groups based on their T1 performance

	content	language
<b>Low achievers</b> (ARC11, DRI04, ICK01, IMJ07, OVD11, USN05, ZEA11)	+0.45 (= 1.78)	+0.72 (= 1.54)
<b>Mid achievers</b> (EBF05, HRG10, LED08, NGS01, UKV05, WAS08)	-0.14 (= 1.76)	+0.27 (= 1.56)
<b>High achievers</b> (ABS04, AKM12, APK08, ARM03, EOD03, UCQ07)	-0.38 (= 1.93)	-0.14 (= 1.88)



shown a similar if not more extreme trend. The fact that the students with highest overall scores on the first test actually lost points on the language scale is even more puzzling.

Before interpreting these results, it should be kept in mind that sample sizes of the three achievement groups are very small, rendering tests of difference (e.g., *t*-tests, Wilcoxon signed-rank tests) useless in this context. Additionally, the grouping is based on the first task performance, which only looks at linguistic features and historical competences that were relevant for this study and further just represents one point in time. Their T1 results might have been affected by several factors, such as decreasing motivation due to the imminent start of the Easter break. In fact, the T2 productions by the students who scored highest on the pre-intervention task were shorter by 78 words on average, indicating that they put in less effort than for the first written task. Low and mid achievers, on the other hand, wrote marginally more in the second sitting (+8 and +6 words respectively). In short, this way of grouping might not be the most reliable approach.

Thus, it might make sense to differentiate the achievement levels on a different basis, such as the teacher's grading of their performance of the previous year, and compare trends. Although we cannot know how these grades were calculated and which aspects they considered, such a grouping might provide a more stable allocation of achievement levels that could also be more relevant for the participants' school experience. When splitting the groups based on their previous English and history grades, the trends described above become less extreme (see Table 14). The

gains of the low-achieving group are now smaller (compared to the gains of students with low T1 scores). The group with average grades made some progress in terms of language and only lost marginally in terms of content. The content results of the group with the best grades also only decreased slightly, while their language results show some limited improvements. In other words, the results based on a different definition of achievement groups

Table 14. Differences T2-T1, divided into achievement groups based on English and history grades<sup>96</sup>

	content	language
<b>Low achievers</b> (ARC11, ICK01, IMJ07, OVD11, HRG10)	+0.23 (= 1.67)	+0.56 (= 1.43)
<b>Mid achievers</b> (DRI04, NGS01, USN05, APK08, ABS04, NGS01, ZEA11)	-0.09 (= 1.73)	+0.40 (= 1.61)
<b>High achievers</b> (AKM12, ARM03, EBF05, EOD03, LED08, UCQ07, UKV05, WAS01)	-0.06 (= 1.99)	+0.07 (= 1.83)

are similar to the outcomes described above, namely that low-achieving students benefitted more from this approach than their peers but to a less extreme degree than a differentiation based on their T1 performance would suggest. At the same time, the results presented above do not indicate that mid and high achievers, no matter which definition, were left behind, as any losses were only small, and differences of results were somewhat equalized when using a more general basis for

<sup>96</sup> The groupings are not evenly distributed as their grades were not evenly distributed either.

differentiation such as annual grades. Looking back at the written feedback, mid- and high-achieving students (EBF05, UCQ07, UKV05, ABS04) also did not seem to perceive overextension.

Differentiation based on T1 performance might allow another inference. Since the results of students with low T1 results improved the most, one could argue that the intervention indeed reflected the issues observed in the needs analysis and also helped them successfully improve in (some of) these matters. Obviously, all interpretations relating to these different achievement groups need to be treated cautiously due to the small sample size and obvious drawbacks of each way of defining achievement levels. Differentiating on the basis of T1 performance relies on a singular performance and might thus not be representative of the learners' "true" competence, whereas grouping based on their previous grades is not specific to the focus of the study, includes a plethora of different aspects, and in the end might also not be truly representative of their level of competence.

#### a) History-based rating

As already indicated above, content results did not change considerably from pre- to post-intervention (see Figure 29). In fact, none of these changes are statistically significant. In some domains, however, effect sizes are medium (*target competence*,  $d = 0.63$ ) or small (*systematicity*,  $d = 0.32$ ;

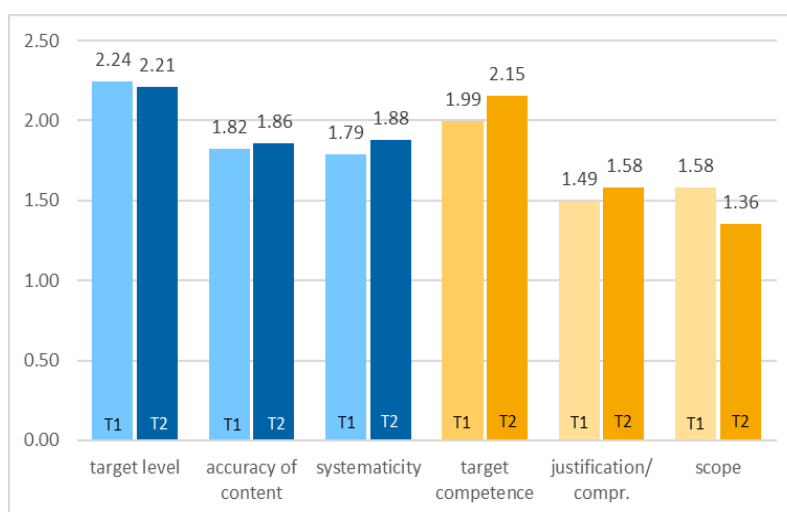


Figure 29. History-based rating: T1 vs. T2, group B, cycle 2

*justification/ comprehensibility*  $d = 0.34$ ; *scope of content*  $r_p = .24$ ). Here, it needs to be noted that scores for *scope of content* actually decreased by 0.22. The second descriptor that declined was *target level*, i.e., performing on the intended level of historical thinking. Yet, this decrease is negligible with only 0.03 points and an effect size of  $d = -0.17$ . These results mean that the content-related issues observed in the needs analysis production (T1) still persist to an almost equal extent. For some learners, of course, some of these initial challenges became less noticeable, while for others new problems arose. ZEA11, for example, often just reproduced facts she remembered from class at T1, thereby losing points because she simply did not display problem-solving skills (*target level*) or a deconstruction of the source (*target competence*) in a comprehensible way (*justification/ comprehensibility*). At T2, however, she seemed to have realized that just presenting declarative knowledge would not suffice. Instead, like in the example below, she clearly articulated her communicative intention (EVALUATE, REPORT) and supported her assessment with an explanation relating to the historical context (EXPLAIN) and with what she saw on the picture (DESCRIBE):

115	RE: {Exploration was not always about exploring, but to find ways how you can export and import goods – The /Handelswege/ were also important to build new relations to an existing country – Many slaves were used for the colonialization – There were many inventions when the explorations began}
Pre-task	
Item 4 ZEA11 (B)	
Post-task (cycle 2)	EV: {I think it depicts truthfully because as I said earlier it was important to have more exports than imports and [EA/ RE: if you want to achieve this, it was necessary to minimize the costs of the transport and production of many goods and products]. [DS: For example, in this picture we see a woman as a colony serving the other woman (the mother country) raw material, to produce cheap goods and then export them.]}

For other students, some problems seem to be connected to limited *scope*, which is a descriptor that started quite low ( $M_{T1} = 1.58$ ,  $SD_{T1} = 0.66$ ) and still lost most points overall, now being the lowest score in the set. Including fewer details and/ or leaving out main points of the task naturally often co-occurred with superficial answers, devoid of acceptable *justification*, which is the second lowest score of the T2 content outcomes. The sample below serves an example:

116	DS: {It is referring to the mercantilism and representing their treatment with minor colonies..}
Item 2	
LOE04 (B)	

LOE04 did not include any details and did not cover the main point of the task, namely exploring potential motives of the artist (*scope* stage = 0). Based on the content, the connection between the source and the answer is somewhat comprehensible, but there is no explicit justification (*justification* stage = 1). In numbers, *scope* strongly correlated with *justification/ comprehensibility*, with  $r_p = .72$ ,  $p = .001$ ,  $n = 19$ . Similar to the pre-intervention results, *justification/ comprehensibility* is the descriptor which throughout showed significant and strong correlation rates to all other content descriptors, with Pearson's  $r$  ranging from  $r_p = .61$ ,  $p = .005$  (target level) to  $r_p = .86$ ,  $p < .001$  (systematicity),  $n = 19$  (see [appendix section II/B/2/ file 3 - correlation analyses](#) for all values). Thus, it could be argued that *justification/ comprehensibility* is a central factor for subject-specific success. It should be kept in mind, however, that to really explore this more thoroughly, one would need to create a regression model determining the factors which impact success in the content subject, and this, in turn, would only make sense with a greater sample size. For the purpose of this exploratory study, it suffices to say that *justification/ comprehensibility* is an area that appears central for content-related success while also posing a problem for the students of group B even after the intervention.

Another issue that appears to be very prominent in this group is the learners' failure to engage more explicitly either with the source or the context. In many cases, it is not quite clear whether the students reported something about the context or whether they described or analysed the source, leading to low history ratings:

117	DS <sub>(RE)</sub> ?: {The queen had all rights. If she wants to gold and silver, she will get it. And so on.}
Item 3	
HRG10(B)	

Item 3 required the learners to elaborate on the connection between the source and real-life mercantilism. With this short answer, it is not clear whether HRG10 kept describing the source (item 1) or whether this could be observed in history. While “the queen” would rather point towards a description of the source, the partial use of past tense could indicate a report of the historical context. Either way, his intention is impossible to deduce. The rating for this answer is 0.8 across the descriptors, reflecting that HRG10 missed the main point of the task and that he only touched upon the issues depicted in the source. As the linguistic links are absent, it is unclear if and how he related to the source; thus, one cannot assume that this would be more than reproduction, which equals historical thinking level 1. Even in seemingly more proficient answers where it is clear what the students were referring to, lack of explicitness could cause similar issues:

118	RE/EA: {With absolutism in France there came mercantilism which means that they strived to export a lot of goods while not importing at all. As they had to get their resources from somewhere they exploited their colonies. .}
Item 3	
EOD03(B)	

In the example above, EOD03 only reported why mercantilists exploited the colonies but did not clarify how or if the caricature relates to this statement, leaving it up to the reader to make this connection. Similar issues were observed in the needs analysis, and it seems that after the intervention, learners such as EOD03 still struggled with verbalizing how reports or descriptions relate to one another, which would be necessary for many analytic tasks, including item 3.

#### b) Linguistic rating

Table 15 shows group B's development from pre- (T1) to post-intervention (T2) written production in terms of language as measured with the linguistic rubric used in this study. All but one of the dimensions improved from T1 to T2, yet only the change of the descriptor with the lowest T1 score is statistically significant, which is *linking in terms of form* (with  $T = 42.0$ ,  $p = .018$ ,  $r_w = .38$ ,  $Mdn_{T1} = 1.0$ ,  $Mdn_{T2} = 2.0$ ). Comparing the T2 results to the observations regarding the students' output gained through the worksheets on Louis XIV, a similar pattern emerges: Cohesive devices were used more often and correctly than nominalised phrases, which in turn were more salient than hedges, reflecting the results of T2.

Table 15. Linguistic rating: T1 vs. T2, group B, cycle 2

	Means		
	T1	T2	T2-T1
<i>CDFs: choice</i>	1.95	2.00	0.05
<i>CDFs: composition</i>	1.44	1.44	0.00
<i>linking/ function</i>	1.21	1.74	0.53
<i>linking/ form</i>	1.05	1.63	0.58
<i>hedging</i>	1.11	1.42	0.31
<i>nominalisation</i>	1.42	1.63	0.21

Looking at the *linking* results of both T1 and T2 text productions, one can see that the mode of both linking descriptors shifted from level 1 to 2 (see appendix section II/B/2/ [file 1 – NA statistics](#) & [file 2 – EV statistics](#) for more information). So, in most student productions, “ideas are generally linked appropriately in relation to CDF type, and linking is linguistically accurate” (= definition of level 2). The number of zero levels decreased from four to one in terms of *function* and from five to two in terms of *form*. Conversely, the number of level-3 ratings grew from two to

three in terms of *function* and from zero to two in terms of *form*. Four students even managed to jump two or even three levels in both linking dimensions. For example, ZEA11's T1 production was rated as level 0 for both linking descriptors. In the second sitting, these numbers changed to level 3 (*function*) and level 2 (*form*), respectively. Extract 115 on page 245 illustrates this development. In her T1 production, she did not use any linking devices, whereas post-intervention, she used several linking devices in appropriate places, in a mostly linguistically accurate way. Similar to the effect size for the difference in terms of *form* mentioned above, the difference for *linking in terms of function* also presents a medium effect with  $r_w = .30$ .

Turning to the learners' use of CDFs, not much change can be reported, with a negligible growth of 0.05 for *choice of CDF type* and no change for *composition of CDF types*. Still, choosing an appropriate CDF type remained the highest value in the set and became the only descriptor with an average of at least 2.0. The group's score for *composition of CDF types*, on the other hand, was the only linguistic area without improvement and, overall, the second lowest value of T2. For many, the issue of creating a CDF composition that supports their overall communicative intention remained an issue. Similar to the results of the needs analysis, in some answers, it was difficult to make out what their main point was:

119	[EV: I do think it depicts truthfully, [RE: since it was/is important that the exports are important to be bigger than the import.][EV: Still, imports are necessary as well, since not every country can have the same resources as the others][EO: so if you want all of the opportunities it's obvious that you're easy to be impressed just by the idea of having "everything".]
Item 4	
ARM03 (B)	

In the example above, apart from issues relating to the accuracy and relevance of the content, the contemplations at the end of ARM03's answer do not seem to sustain her first evaluation concerning the validity of the source, as she drifts off. Additionally, all CDF episodes following her first assessment deal with the importance of imports and exports from various angles, although it is not very clear how these parts relate to the authenticity of the source. Still, this issue of confusing CDF compositions resulting in unclear overall communicative intentions seems somewhat less pronounced in the post-intervention texts. Instead, for some, the reason for unclear CDF-composition often lies in the assemblance of basic-CDF elements where none is easily identifiable as the CDF episode, i.e., overall communicative intention. Such an answer often appears somewhat indirect in tackling the main point of the task, as the following example illustrates:

120	RE <sub>(EV)</sub> : {Because the European and other countries all over the world imports raw materials, gold silver, etc. from this "colonies" for example from Africa, [CA: but nowadays we have to pay for this materials but they are still from that countries.]}
Item 5	
APK08 (B)	

In APK08's statement, it is not entirely clear whether this report on current exploitation and his ensuing comparison between past and current practices can actually be considered as an evaluation of the relevance of the source for today's world. Additionally, the comparison is partial since he only implied that import/export in past and present were similar except for one aspect. Overall, APK08 did not make it explicit enough how these elements play a role in the source's

relevance in the 21<sup>st</sup> century, which is further complicated by the answer's syntactic structure, essentially being a string of ill-linked subordinated clauses.

In the pre-intervention task, *hedging* was, alongside cohesiveness, one of the weakest areas. After the intervention, a positive trend could be observed, presenting a small effect of  $r_w = .21$  but no statistical significance. What can be said, however, is that the number of zero levels went down considerably, from five students to only two students who did not employ any strategies to hedge their claims. While the mode for *hedging* remains on level 1, the number of students at level 2 changes quite noticeably from one student at T1 to six students at T2. In general, however, the changes observed are not drastic, with only one student (ICK01) improving by two levels:

121	EA/EO: {So that we can laugh about it because no one in the western world <b>would</b> take this serious because in this part of the earth we just find it funny. He <b>might have produced</b> this cartoon because this is how people lived years before us and that he can show us with the picture that we are all “slaves” for our politicians.}
Item 5	
ICK01 (B)	

ICK01, who only employed one hedging device in her whole T1 production, used two qualifying forms in this item alone and six in total. As an aside, this sample also shows that she (as well as many of her peers) still struggled with exploring motives in their historical context and only considered a 21<sup>st</sup>-century audience (presentism). Finally, it should be mentioned that the issue of inappropriate boosting, as opposed to hedging, was less apparent in the post-intervention performances than in the pre-intervention texts.

Lastly, some modest positive changes in the learners' use of *nominalisation* can be reported, with an effect size of  $r_w = .29$  and six students improving their *nominalisation* ratings. However, only one student changed by two levels, namely LED08. The ratings of ten students remained unchanged and three ratings even decreased. The rather considerable number of unchanged ratings, nonetheless, does not necessarily mean that the learners' use of nominalised phrases did not change from a qualitative point of view. What can be observed in many written productions here is that learners, by and large, used a more nominal style than in their pre-intervention texts, but they also tended make more mistakes, indicating level 1 of *nominalisation* (“some evidence of nominalisation, but may be partly used inappropriately”), which was given to nine students. Examples would be “absolutism king”, “colonializations” for the general phenomenon, “the strengthen”, “that results to the wealthy of Mother country” or “their wealthy and status”. In the pre-intervention evaluation, level 1 was assigned nine times too, yet only two instances of linguistically inaccurate nominalisation were marked in the data, suggesting that the main reason for level-1 ratings was infrequent use of nominalised phrases. This would suggest that the linguistic support on nominalisation might have risen the students' awareness but not necessarily their competence. Interestingly, when looking at the lesson transcripts, the students' issue of confounding nouns, verbs, or adjectives of the same word family can also be seen in the teacher's own oral production:

**Extract 122 and 123, lesson 3, cycle 2:**

- 122 TB2: Perfect. **Weak**s and strengths from your opponents, **competitives**. Perfect examples, ähm, where are your opponents or **competitives**?
- 123 TB2: [...] to show how **strength** you are

However, this is only to be understood as a tentative observation. The data available is unfit and too limited to explain a potential link between student and teacher production.

c) Summary and implications for the design

In group B, the observed changes in the students' writing from pre- to post-intervention are limited. Especially in terms of content, hardly any substantial change could be observed, suggesting that the learners' performance of historical skills remained mostly unaffected by the intervention, unlike in group A, who performed better in terms of content after the first intervention. When it comes to linguistic skills as measured by the rubric used for this study, the students of group B could show some improvements. Just like in group A, cohesiveness improved most considerably (and, in part, statistically significantly), followed by use of hedging and nominalisation. Use of CDFs, however, virtually stayed the same. Here, it should be noted that Group B did not only experience a CDF-based intervention, but in essence, a CLIL intervention more generally, having hardly experienced BE prior to the project. From this perspective, group B's zero effect in terms of content ties in with the often reported observation that content learning remains unaffected by CLIL, while language learning seems to benefit (e.g., Dallinger et al., 2016; Dalton-Puffer, 2008; Pérez Cañado, 2012; San Isidro & Lasagabaster, 2019).

Splitting up the group into different achievement levels has shown that low-achieving students greatly benefitted from the approach, whereas mid-achieving and high-achieving students only improved marginally or even lost points, depending on the definition of achievement groups. Differentiation on the basis of T1 outcomes shows rather considerable differences of developments. Those learners with the lowest T1 scores greatly improved both in terms of content and language, whereas students with average scores could only gain points in terms of language, and the ratings of students with high T1 scores decreased in both domains. As those with the lowest T1 scores improved the most, it could be argued that the intervention indeed reflected the issues observed in the needs analysis and helped the learners improve in those areas.

Grouping the learners on the basis of previous grades, however, reduces the differences between the developments of the achievement groups. This way, on average, the CDF-ratings of all groups improved, whereas the content scores of the mid- and high-achieving students slightly decreased. Comparing these results to the different developments on the basis of T1 performance indicates that no clear trend can be observed, potentially as a result of small sample sizes that are sensitive to individual trajectories. Students with low grades, however, improved the most in both domains, similar to students with low T1 ratings, indicating a more stable trend, which also corresponds to the results by Lo and Jeong (2018), who examined the impact of a genre-based approach on the

academic literacy of secondary history learners in Hong Kong. Measuring only their academic literacy, they report that weaker students benefitted more than stronger students.

Based on the results described above, two main implications for the design arise, which to some degree match with previous insights:

First of all, the next cycle should consider content aspects to a greater degree to make sure that students also improve their performance of historical competences. As this study intends to approach bilingual learning from a content-and-language-integrative angle, scaffolding and support boxes should engage more clearly and more deeply with the relationship between certain linguistic choices and subject-specific content and skills. While this reflects group A's wish to aim for more history-focused linguistic support, it contradicts group B's wish for more general support, which might be explained by their limited experience with BE. As cycle 3 is set again in context A, the following unit tries to exploit the idea of subject-specific linguistic support more straightforwardly. It should be kept in mind that in a different setting, i.e., where experience with BE is limited, one might opt for a less specialized approach initially, which then becomes more and more subject-specific and in-depth.

In the context of cycle 3, the new unit should focus more on explicitness, i.e., clear verbal expression of how different elements relate to each other. Such a so-called *visible pedagogy* (see Bernstein, 1999) explicates cognitive steps and demonstrates the language required for these processes. This way, the learners might be able to move from mere reproduction to genuine and recognizable analysis, reflection, or problem-solving, increasing their thinking level ratings. This seems to play a central role for the students' real-life academic success since many Austrian history teachers grade with such a three-step thinking level matrix in mind, reflecting the grading system of the final exam (Kühberger, 2011). A central element of these thinking skills is comprehensibility and justification, which would also greatly benefit from a clear verbalisation of their thought processes. In turn, a clear verbalisation of cognition, it is assumed, requires a logical composition of CDF types that are appropriately linked.

Secondly, CLIL instruction and scaffolding needs to be more differentiated in order to cater for different ability levels (see, e.g., Donato, 2016; Lialikhova, 2019). It appears that mid and high achievers were not supported the way they would have needed at this point. Looking at content learning, it seems that they could not keep up their relatively high levels when dealing with content in English as compared to their T1 performance, where they had acquired the assessed content in German. Thus, one needs to keep in mind that even for high achievers, this approach could pose a challenge, which T<sub>B2</sub> stressed in his interview too. As a consequence, gifted learners might also benefit from support measures, but they would need different tasks and scaffolding to reach even higher. As a consequence, the materials need to consider different ability levels to a more substantial degree and not just for the sake of counteracting diverging work paces, as mentioned so often in the context of group A, but for the sake of providing the best care for all types of learners.



### 7.4.3 Cycle 3: the Industrial Revolution (school A)

Directly after the implementation of the unit on the Industrial Revolution (IR), students of group A were interviewed to evaluate the intervention from their perspective. To be able to follow up on previous insights and student beliefs, everyone participating in preceding interviews and present on the day of this interview was invited. As mentioned before, due to student absences on particular days, the sampling of these group interviews varied. Additionally, to avoid an imbalance of achievement levels and types of learners, further students were invited too, totalling in seven interviewees, namely ARJ01, ATC04, ETS12, EVS03, HIP11, IJT12, and ORH09. In order to also follow up on students that were not present on this day and to be able to get a more comprehensive picture of the students' evaluation, the brief feedback survey used in cycle 2 was also distributed to the students of group A who did not participate in the final interview.

One week after the final lesson of the intervention, the students of group A completed another written competency-based task to monitor and assess their development. This time, the students were asked to analyse a photograph showing the effects of urbanisation during the IR (see [appendix section I/B/ file 1](#)). To be able to discuss and contextualize preliminary results of the written task and the interview with the students, the teacher interview took place one month later.

#### 7.4.3.1 Retrospective interviews with students and teacher

##### a) Experiences with the intervention

After the second round of intervention, both students and teacher reported a predominately positive experience, which Figure 30 illustrates. Unlike in cycle 1, both teacher and students perceived the linguistic

pedagogical practices as mostly positive, indicated by thick green lines and thin red lines, reflecting the frequency of co-occurring codes. The teacher especially was very satisfied with the implementation of this unit, describing the

experience as pleasant and easy. She could not think of any issues, including problems faced during the first intervention. She felt that the two main issues of cycle 1 had been resolved. To her, the language focus was less salient than during cycle 1 but still more noticeably than in her usual lessons. In the future, she would like to continue with the balance practised in cycle 3. In her very first interview, T<sub>A</sub> said that when it came to the relations of content and language teaching, she

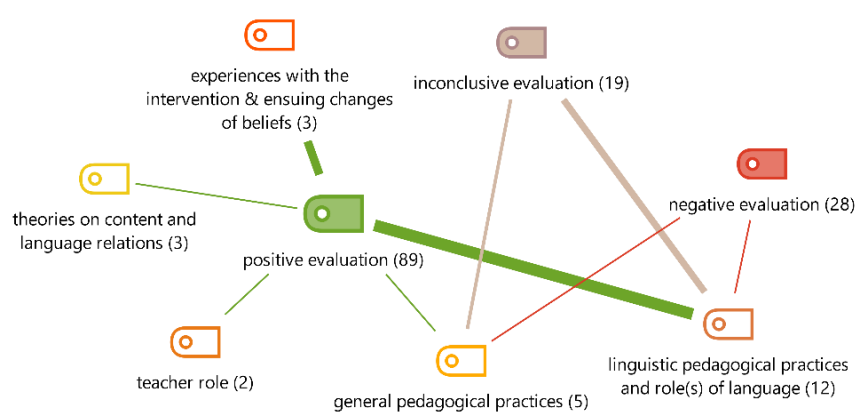


Figure 30. Code co-occurrence model: experience of students and teacher, group A, cycle 3

did not know what the future would bring. Confronting her again with this quote (see p. 134) and asking for her thoughts on this matter, she stated:

124	English translation	Original quote
T <sub>A</sub>	Well, <b>continuing in this direction for sure</b> , because [...] I always come back to <b>performative verbs</b> [...], because we want them to do specific things, and not just in English or history but in all subjects. And I realized that they've got <b>huge problems</b> with that because <b>only very few subjects show them or tell them</b> what they need to do, but it is expected [...] So far, I've only done this in English of course, but I'll also pay attention now in history for sure.	<i>Ja, nein auf alle Fälle in der Richtung weitermachen, weil [...] ich komm immer wieder auf die Operatoren zurück, [...], weil wir ja wollen, dass sie bestimmte Dinge tun und zwar nicht nur jetzt in Englisch und in Geschichte, sondern eigentlich in allen Fächern. Und ich drauf gekommen bin, dass sie damit massive Probleme haben, weil das in den wenigsten Fächern eigentlich wird ihnen gezeigt oder gesagt, wie sie das machen sollen, sondern es wird erwartet. [...] Ich hab das bis jetzt nur immer in Englisch natürlich gemacht, aber werd halt auch in Geschichte wirklich drauf aufpassen.</i>

Having performative verbs and by extension testing in mind, she realized that the learners really struggled with clearly expressing different cognitive operations, and thus far these learners had not been supported well in this regard. Therefore, she intended to continue in this direction, not just in English classes but also when teaching history, both in bilingual and mainstream classes. One reason for cycle 3 working out better than cycle 1 was also the process of familiarization with the project. Apart from improving the approach, T<sub>A</sub> argued that she experienced cycle 3 in a much better way because she and the learners had gotten used to the method:

125	English translation	Original quote
T <sub>A</sub>	<b>I'm getting used to the method</b> (laughs). [I] can image continuing like that to some degree. [...]	<i>Ich gewöhn mich langsam an diese Methode (lacht). [Ich] kann mir das auch vorstellen eben weiter so zu machen, in einem bestimmten Ausmaß.</i>
	It just worked well, but, I think, because the students had gotten used to it by now. Well, <b>they just know what we want from them.</b>	<i>Es hat einfach gut funktioniert, aber ich glaube auch eben, weil's jetzt eben die Schüler jetzt schon dran gewohnt haben. Also sie wissen einfach, was wir von ihnen wollen.</i>

A similar effect of adjustment could also be observed in the student interview. While in their previous interview, the students of group A repeatedly complained about the strong language focus and the teacher's insistence on using the phrases, they now seemed more satisfied, confirming that content and language appeared more balanced. Looking at the lesson transcripts, in contrast to cycle 1, T<sub>A</sub> only rarely commented on the students' use of language in whole-class discussions in cycle 3. Yet, during working phases, T<sub>A</sub> often provided feedback concerning their use of language and sometimes also pointed towards the boxes. Interestingly, however, although T<sub>A</sub> did not openly insist on the use of phrases anymore and just sometimes mentioned the existence of these boxes, the learners started to consider these boxes in their own language use more generally without the teacher having to enforce it. This could be especially observed during the working phases, as many students asked their teacher or their peers for linguistic feedback and support, as indicated by the following example:

**Extract 126, lesson 1, cycle 3:**

- 1 UYA0 [to ETS12]: Wenn ich das da verwende, dann muss ich das da auch oder? **[If I use that one, do I have to use this one too, right?]**
- 2 ETS12: Mhm [negating] Nein, du kannst einfach nur, du musst zuerst einmal ausdenken, was für einen Satz du schreiben möchtest und dann schauen, ob du das mit diesen Sachen verbinden kannst. [No, you can just, you have to first think about what kind of sentence you want to write, and then you check whether you can link these things].
- 3 UYA06: So do I need this one?
- 4 ETS12: Ja, was willst du damit sagen? [Well, what do you want to say?] [...]
- 5 TA [to ETS12 and UYA06]: **Is there a problem? Or a question?** [...]
- 6 UYA06: I don't know how to use this X
- 7 TA: How do you use this phr-, if it's correct?  
[...] ((TA reads UYA06's answer))
- 8 TA: Mhm [affirmative], das geben wir weg [You can cut this one]. Manual labour and draft-animal agriculture, so and you don't need the brackets. Punkt [full stop]. Okay, the sentence is far too long. Okay.
- 9 ETS12: Ich würd sagen, damit kannst du einen neuen Satz anfangen **[I'd say you could start a new sentence with this one]. ((pointing to a phrase))**
- 10 TA: Genau [Exactly]. It depends how you want to continue. Yeah, genau [exactly], but it's okay. As a result blablabla there were, okay.

UYA06 wanted to include some of the phrases but did not really know how and thus asked ETS12 for help (line 1). T<sub>A</sub> noticed this and offered to assist (line 5). After reading her sentence, T<sub>A</sub> provided feedback (8) and ETS12 made some further suggestions (9), helping UYA06 include the phrases in an appropriate way. Other examples found in the transcripts include asking about the appropriateness of certain linking devices, how one could leave out the agent (passive voice), or generally whether a certain formulation was clear enough. Furthermore, a shift in awareness can also be seen in the students' use of language. When comparing answers in class, most learners made use of the phrases provided and seemed to consider the language boxes, as the examples below illustrate:

**Extract 127, lesson 1, cycle 3:**

- 1 AVS07: There were many radical changes, **for instance, the increase of production. Furthermore, manual labour was replaced by machine-based manufacturing** in agriculture. **Moreover**, there was an **improvement** of the roads and railways, which **led to the trade expansion**, and in the 19th century, the Industrial Revolution started spreading towards Western Europe and North America, **resulting in industrialisation**, industrialisation all over the word. [...]
- 2 TA: In paragraph three you had to report different views and the exact time frame of the Industrial Revolution, and you had to give an explanation. AVS07, what's your explanation?
- 3 AVS07: Historians argue about the period of time covered by the Industrial Revolution. Some **assert that** it started in 1780 in Britain and ended between 1830 and 1840. **But it's claimed** that it started between 1750 and 1830

**Extract 128, lesson 3, cycle 3:**

- 1 SAA03: I **could imagine** that the artist drew the picture because he wanted to create **awareness of London's poverty**. [...]
- 2 ICM01: I think that he **might have drawn** it that dark to **convey the situation** or the feelings the people had.

In lesson 1, AVS07's answers fully considered the language boxes and the phrases provided. Her first contribution included many phrases for structuring her answers as well as phrases for

expressing cause-and-effect. In the next task, AVS07 used different reporting verbs, taking into account the language box on reporting views and opinions accompanying this task. In the second extract, the learners used modal verbs to highlight that they were only speculating about an artist and the style of an image, as suggested in the language box on their worksheet. Additionally, all three students in the extracts above used a very nominal and formal style. It seems that the teacher did not really have to insist on considering the boxes, as the students were doing this of their own accord. She sometimes also praised students for using the phrases, as illustrated in Extract 129:

**Extract 129, lesson 2, cycle 3:**

- 1 HIP11: Ähm according to this graph
- 2 TA: = **Good**. ((HIP11 looking up)) Good.
- 3 HIP11: Aso.
- 4 TA: **You are finally using phrases I want you to use.** [...] Was heißt denn according to in German? [What does according to mean in German?]
- 5 Sxx: Laut.
- 6 TA: Laut. Here we go. Once again from the beginning, HIP11, I'm sorry.

Here, HIP11, a student who appeared somewhat resistant to the language support, used phrases provided. The teacher's praise even seemed to surprise him. As a follow-up, T<sub>A</sub> briefly asked for a translation of the phrase and then let HIP11 continue. Another form of reacting to the students' language use was corrective feedback in terms of the linguistic features in focus. The following extract serves as an example:

**Extract 130, lesson 3, cycle 3:**

- 1 EVA02: Ähm, since there isn't or, ähm, the room seems to me as if it is, äh, dark because there isn't much light coming in and, äh, it's quiet because everyone is asleep.
- 2 TA: Okay, **the room seems dark?**
- 3 EVA02: Ja. Is dark.
- 4 TA: Ja, I would also say it is dark because or since, ja, as the reasons you've given.

In this extract, EVA02 used a hedging phrase in an inappropriate context. T<sub>A</sub> repeated this part with a rising intonation, which EVA02 correctly interpreted as a request to rephrase.

Given this shift of awareness in both teacher and students, the teacher's assumption that the second intervention worked well because the learners knew what was expected from them seems to hold. Conversely, one can also argue that the implementation worked better the second time around, as the design team, including the teacher, also knew and were mindful to what the students expected, creating a balance of content and language considerations that enabled a better working atmosphere.

Other issues mentioned in the previous interview were diverging work paces and a high level of exhaustion. In the second post-intervention interview, quick learners, like ORH09 or ARJ01, reported that they completed fast-track activities and thus did not experience any undesirable waiting times. Concerning the level of intensity, the learners did not mention anything in this regard, indicating that this was not perceived as particularly negative (or positive). Thus, also these problems appear to have been solved in this cycle.

## b) Evaluation of the intervention

Similar to the participants' rather positive experiences, their evaluation of the materials shows high satisfaction levels. The code co-occurrence models in the following illustrate which aspects were positively, negatively, or inconclusively assessed in the interviews with the students and the short student survey, pooling the views of seven student interviewees and eleven respondents (Figure 31), and in the interview with the teacher (Figure 32). While the teacher's evaluation was overwhelmingly positive, with 28 positive, three negative, and only two inconclusive codings, the students' assessment was considerably more varied. Given their number, it is not surprising that views were somewhat mixed, but nonetheless a trend towards a positive evaluation can be observed, with positive codings (61) clearly outnumbering the negative ones (25), which is in stark contrast to cycle 1. Even when subtracting the survey results to better compare cycle 1 and cycle 3 results, the picture is still mostly favourable, with 28 positive codings compared to 12 negatively coded remarks.

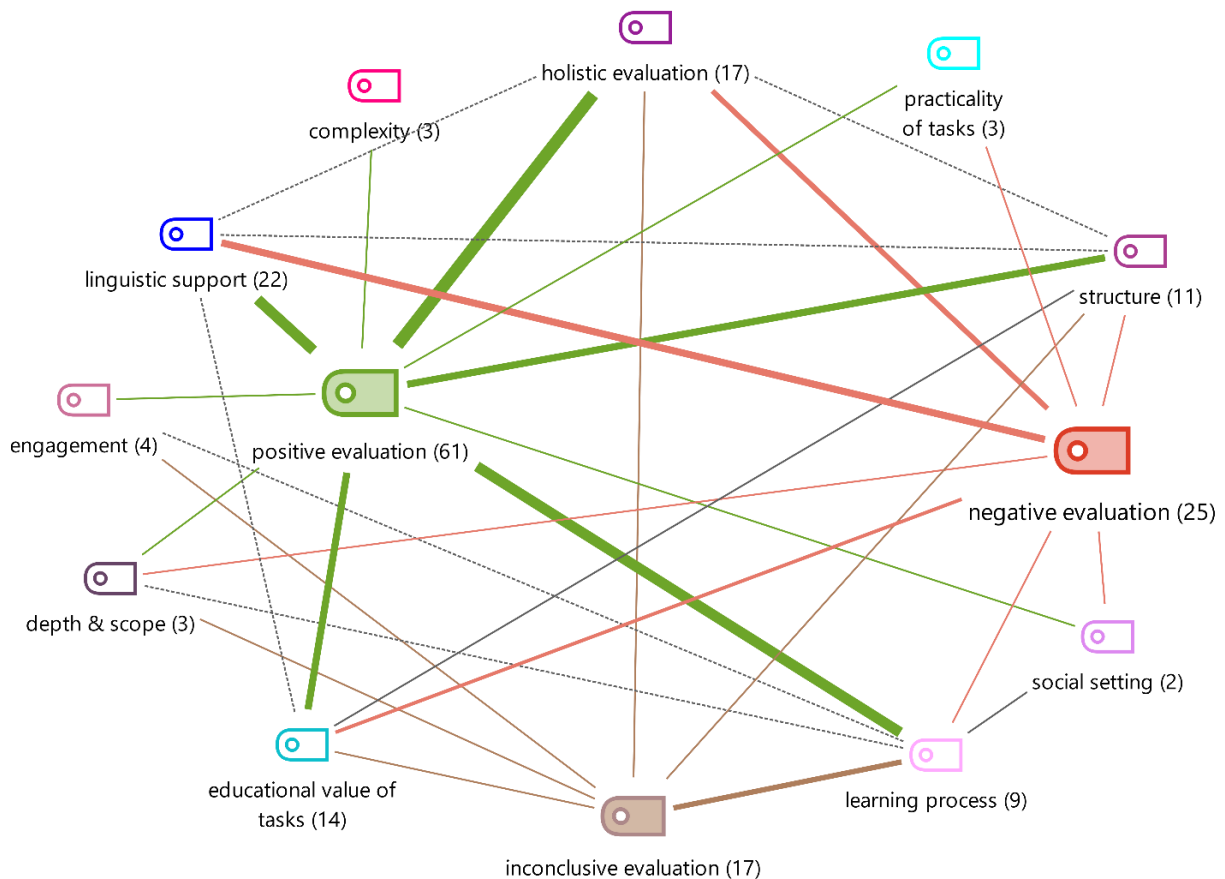


Figure 31. Code co-occurrence model: cycle 3 evaluation by the students of group A

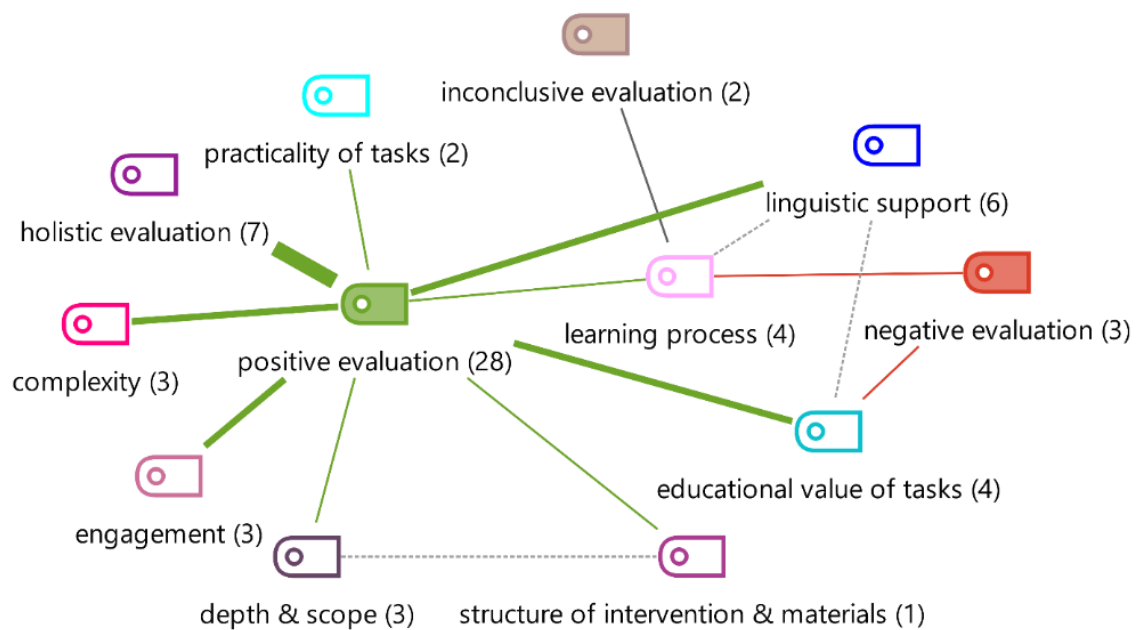


Figure 32. Code co-occurrence model: cycle 3 evaluation by TA

Starting with **engagement**, the teacher highlighted higher levels of participation as one of the most memorable aspects of cycle 3:

131	English translation	Original quote
TA	R: Is there anything that stuck in your mind as particularly positive? TA: <b>Eh, yes, just that more students participated.</b> Well, more than last time.	R: Gibt's irgendwas, was dir besonders positiv in Erinnerung geblieben ist dieses Mal TA: Äh, ja das einfach mehrere Schüler mitgearbeitet haben. Also mehr als beim ersten Mal.

In her cycle 1 retrospective interview, TA already stated that students had participated more in cycle 1 than in her usual classes, while in her pre-intervention interview, TA lamented the lack of motivation and low levels of engagement as main issues. Thus, one could argue that the intervention addressed this 'local' problem successfully. The students mostly confirmed higher levels of participation. More quiet students, like ATC04, added that they were always cognitively engaged even if they did not say that much in class:

132	English translation	Original quote
ATC04	<b>I participated more, well I also had to think more for myself, yet I didn't talk much</b> , but that's rather [my personality].	Ich hab mehr mitgearbeitet beziehungsweise musste ich auch mehr mitdenken, aber ich hab trotzdem nicht viel geredet, das ist eher so [meine Persönlichkeit].

Still, the more proficient and usually very active students, like ARJ01 or ORH09, were not sure whether they participated to the same degree or less, as this depended on whether the teacher actually called upon them, implying that normally, not that many other students would raise their hands to contribute to whole-class interaction, as already stated in a previous interview.

**Complexity** was also rated favourably by both learners and teacher, stating that the complexity level of input and tasks was appropriate. Some learners mentioned that the texts, especially the

historic quotes, were challenging but doable with the help of the glossary. The teacher also added that increasing familiarity with the task type and the methodological steps, i.e., how to approach such sources, made it more accessible.

Unlike other interviews, nothing was said concerning **social setting**, indicating that they did not notice anything particularly good or bad in this regard. Only in the written feedback, two students voiced their opinion on this matter. One respondent appreciated that there was individual and group work, which helped them understand, while another student wished for more interactive tasks. Concerning the **learning process**, T<sub>A</sub> and all students agreed that the students learned something in terms of content. Some learners added that, in normal lessons, they would have learned just as much or even more because the pace was faster, covering more topics, in traditional lessons. Yet, the majority felt that they made considerable progress in terms of historical content. In the written feedback, they were asked specifically if they had learned something and why:

133	English translation	Original quote
several students	- <b>Everything was explained in detail.</b> I learned something and I hope I won't forget it	- <i>Alles war ausführlich erklärt. Ich habe was dazu gelernt &amp; hoffe das ich es nicht vergesse</i>
WF	- Yes, <b>I paid attention</b> - Yes, because we discussed more often - Yes, <b>it helped me express in an interesting way</b> - Yes, it was very <b>easy to remember these things</b> - <b>Well-structured</b> handouts; <b>group and individual work</b> for better comprehension	- <i>Ja, ich hab aufgepasst</i> - <i>Ja, weil wir öfters besprochen haben</i> - <i>Ja, es hat mir geholfen intressant zu formulieren.</i> - <i>Ja, es war sehr einfach mir die Sachen zu merken</i> - <i>übersichtliche Handouts;</i> <i>Gruppen/Einzelarbeiten zum besseren Verständnis</i>

All students answered “yes”, and some added reasons as presented above. This matched their teacher’s perception, who reported that in terms of content, the students definitely improved, as this could also be seen in their grades. When it comes to linguistic outcomes, opinions were more mixed. In the written feedback, language was only mentioned once (“it helped me express in an interesting way”), while in the interview, five out of seven felt their language improved somewhat. Similarly, T<sub>A</sub>, being also their English teacher, did not perceive any substantial change in their use of English, arguing that the intervention was too short to lead to persistent improvements. She still appreciated that some students felt that they improved linguistically too, as this could boost their confidence.

Turning to **scope and depth** of the intervention, the student and teacher interviewees agreed that the intervention had an appropriate degree of detail and depth. Only one respondent of the survey wished for a faster pace and less depth to be able to cover a greater number of topics. The **structure** of the materials and the intervention was also mostly met with positive feedback by students and teacher alike. Most students valued the structure of the lessons and the handouts. Fast-learning students pointed out the usefulness of optional tasks. ORH09, for example,

appreciated not having to wait for others and mentioned that some of the language boxes accompanying the fast-track exercises actually provided her with new and useful input:

134	English translation	Original quote
ORH09	We [= ARJ01 and ORH09] highlighted a box, the one with strong verbs and neutral verbs and stuff like that; we found this one good because, of course, one knows how to express those things, well, that someone says something or something like that. But it is good to have it that clearly summarized [...] if it is strong or not that strong <b>because one might not yet know that.</b>	<i>Wir haben eine Box angestrichen, das war die mit den strong verbs and neutral verbs und solche Sachen, ähm, die fanden wir halt gut, weil man, also klar man weiß, wie man Dinge ausdrückt, so, dass jemand das sagt oder so. Aber es ist halt gut, wenn man's so übersichtlich hat [...] ob das jetzt stark ist oder nicht so stark, weil das weiß man vielleicht noch nicht so.</i>

So even though she felt that she knew how to report someone's views, she appreciated having a more in-depth explicit comparison of different reporting verbs.

In connection to the clear structure, the students also welcomed the precise labels and instructions, thus rating the **practicality** of the tasks as positive. This corresponds to T<sub>A</sub>'s observation that during working phases, she hardly received any questions on what to do. The only aspect criticized in regard to practicality was the printing quality of the visual source.

Turning to the **education value**, like speculated after cycle 1, a certain degree of novelty appeared to be necessary for students to perceive tasks and linguistic support as valuable. Of course, what can be considered new to students greatly depends on the students' ability levels and previous experiences. Additionally, perceived novelty, obviously, was not the only factor for the students' awareness of educational value. For example, all student interviewees marked part 2, i.e., describing graphs of population development, mostly in green, highlighting the usefulness of the linguistic support:

135	English translation	Original quote
ETS12	For example, when we had to describe those diagrams, the linking devices, etc., really helped, I think.	<i>zum Beispiel als wir diese Diagramme beschreiben mussten, dass sie glaub ich auch geholfen haben mit den linkern und so.</i>

T<sub>A</sub> later clarified that this was nothing new to the students, as she had just recently taught language for graph descriptions, including linking devices, in English class. Apparently, however, the students still appreciated the revision, now in the context of history education, as the green markings in Figure 33 illustrate:

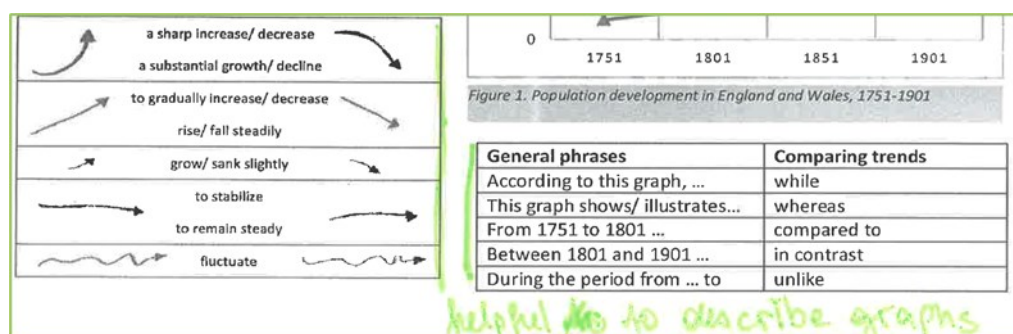


Figure 33. ETS12's markings on the handout



In general, with seven different voices, several aspects of the intervention were marked differently, with comments ranging from ‘useful’ to ‘unnecessary’ for the same task or language box. In the student interview, most tasks were coloured in green by all students, with only a few exceptions, namely the visual source description, speculating about motives for producing such artwork, and the related impulse question. Two students argued that describing visual sources was pointless:

136	English translation	Original quote
ARJ01 & ORH09	<p>ARJ01: Let’s put it like that, I hate every task involving image description since <b>everybody can see what’s in the picture</b>.</p> <p>ORH09: = Yes [...] For us, it’s just logical actually [...] <b>everybody can interpret a picture any way he wants</b>, which is not really the case with texts.</p> <p>R: Well, but you always need to provide good reasons. That’s the skill.</p> <p>ORH09: = But I can.</p>	<p><i>ARJ01: Sagen wir so, ich hasse jeden task, der so Bildbeschreibung ist, weil jeder kann schauen was am Bild ist.</i></p> <p><i>ORH09: = Ja [...] Für uns ist das halt eigentlich logisch [...] jeder kann halt reininterpretieren in ein Bild was er will, was jetzt bei einem Text nicht so der Fall ist.</i></p> <p><i>R: Naja, aber man muss immer gute Gründe geben. Das ist die Kunst.</i></p> <p><i>ORH09: = Kann ich ja.</i></p>

ARJ01 and ORH09 presented some sort of ‘anything goes’ approach when it comes to visual source analysis. For them, description was a logical, straightforward process, and backing up interpretations with good reasons felt easy. T<sub>A</sub>, on the other hand, did not agree, stating that the learners of group A still struggled very much with these methodological skills, even if they did not realize it themselves. As for the impulse question, which aimed at transferring these skills to the present day, one student argued that it did not fit to the content and was thus unnecessary. Additionally, for some, part 4 felt “too much like English class”, as it involved a considerable amount of writing. Additionally, the wording of the prompt, including the word limit, reminded them too much of English homework. At this point, it should be noted that, typically, Austrian history lessons rarely feature any writing. Therefore, some students viewed this part as inappropriate for history class. In the survey responses, most students seemed satisfied with the nature of the tasks and the educational value of the unit. Some explicitly appreciated the variety of input and sources and the creativity of the tasks. One student found “all the summarizing unnecessary”, which, like the cycle 1 interviews, indicates that some students lacked awareness concerning underlying communicative intentions of performative verbs, as everything was considered to be “summarizing”.

One of the biggest differences to the retrospective interview after the first cycle relates to the students’ assessment of the **linguistic support**. After the first intervention, most students expressed negative or, at best, rather critical views, whereas now most of them appeared to value them to some extent. All interviewees found at least some boxes helpful, and some students even marked all boxes in green or specifically highlighted their usefulness, including students who disapproved of the linguistic support in cycle 1, such as ETS12 or HIP11. As mentioned before, high-performing students like ARJ01 or ORH09 only appreciated boxes that presented new information, such as subject-specific linguistic input. Reflecting a hypothesis by T<sub>A</sub>, the students

confirmed that in history lessons, they preferred linguistic support tailored to the subject. ARJ01 and ORH09 further expressed annoyance with boxes that explicitly elaborate on communicative intentions of certain performative verbs, as they felt that they already knew that. Weaker students (in terms of school grades and written task performance), such as IJT12 or ATC04, appreciated all boxes. One explanation would be that weaker students might not have been aware of most of the information presented in these boxes, making it feel ‘new’, while stronger students had already internalized them. As stressed by T<sub>B2</sub> in cycle 2, the purpose of scaffolding is to help students reach higher until a point where such support measures are not necessary anymore. From T<sub>A</sub>’s perspective, these language boxes do help them respond adequately to performative verbs and express themselves more clearly:

137	English translation	Original quote
T <sub>A</sub>	I find it [= linguistic support] <b>extremely important [...] so that they have scaffolds</b> with the language or these boxes you provided. So that they know, ‘okay’, what language or which words can I use for which performative verb. <b>I think that they really get a better feeling for this; I hope at least.</b>	<i>Ich find’s [= sprachliche Unterstützung] enorm wichtig [...] damit sie auch ein Gerüst haben mit den, äh, mit der Sprache oder den Boxes, die du da angegeben hast. Damit sie wissen, okay, welche Sprache oder welche Vokabel kann ich für welchen Operator verwenden. Ich finde, da kriegen sie mehr Gefühl, oder hoffe ich es zumindest.</i>

She further noted that some students in need of support still rejected these language boxes, for instance HIP11. Although HIP11 stated that he liked having phrases as inspiration, he rejected them as soon as he felt pressured into using them. On the handouts, he even marked phrases like “with the help of the boxes” in red as they made him feel obligated to really use them. One can only speculate, but while he might see the benefit of linguistic support, the thought of actually performing them in class could make him feel anxious. Coming back to the needs of more proficient learners, T<sub>A</sub> agreed with T<sub>B2</sub>’s view that such students would just need to learn to ignore support measures below their level.

### c) Summary and implications

Overall, students of group A seemed to approve of the materials of cycle 3, with most students giving them positive overall ratings and pointing out many aspects of the materials they particularly liked, such as the type of tasks and input, subject-specific linguistic support, the level of complexity and engagement, and more. The teacher, who was already quite satisfied with cycle 1, now expressed unreserved approval, asserting that she would continue with this approach, even in mainstream classes. The main issues of cycle 1, namely diverging work paces, overload, unspecific linguistic support, and a too prominent role of linguistic features during classroom talk, could be solved. Optional tasks balanced out different learning speeds, and density of source-analysis tasks was reduced. Linguistic support was more specific to the field of history, and the use of phrases was not in focus when discussing content. Instead, the teacher provided linguistic feedback more subtly during work phases. Interestingly, the students started to call for such support during elaboration phases. Still, it appears that one needs to create a setting distinctive of

the content subject and which does not feel too much like EFL teaching, e.g., when it comes to the formulation of prompts or tasks.

While a number of issues were resolved in cycle 3 with the help of the participants' previous feedback, another factor facilitating approval could be familiarization with the approach. It seems that the teacher and students became accustomed to the new approach, which facilitated better results, both in terms of acceptance by the students and actual learning outcomes, as the following subsection will show. From this, it can be deduced that an introduction of a new method that is considerably different from normal lessons requires the planning of at least two cycles with the same student group to lead to satisfactory results within this particular group. This ties in with recommendations by experienced DBR researchers, such as Eijkelhof (2017), who pointed out that any innovation takes time to get accustomed to and thus advises against control groups in early cycles of design research (see also Edelson, 2006).

In the case of group A, this second round of intervention has shown that incorporating participants' voices in the planning of didactic innovation supports participant approval and ultimately improves the perceived and actual learning outcomes of the students, validating calls for more active student involvement in educational research (e.g., Coyle, 2013, Filice, 2021, or Groundwater-Smith & Mockler, 2016). More specifically, several important factors for the success of this unit could be identified. These include ensuring subject-specific linguistic support for students experienced with BE and treating it like that during the lesson (i.e., always discussing linguistic choices at the backdrop of the discipline), planning for different ability groups, and aiming for a feasible workload. These aspects should be considered when preparing other similar units. At the same time, it needs to be kept in mind that planning for different ability groups was only rudimentarily done in cycle 3 and would need further attention when designing new sets of materials or revising this one. In terms of differentiation between different learning paces and ability groups, one concrete idea for part 4 would be to make the use of the historical sources provided via the material optional and invite faster learners to do their own research. The need for differentiated instruction also includes the linguistic support provided, meaning that different ability levels require different linguistic support, trying to make nobody feel patronized and/or bored, or, on the other side of the coin, overwhelmed and discouraged (see also Tedick & Young, 2018). In a way, one needs to try to create linguistic support that presents new information in relation to the students' respective ability level and what they have already internalized (see also Donato, 2016; Lialikhova, 2019).

### 7.4.3.2 Post-intervention written tasks

Figure 34 illustrates group A's development of written task performance throughout the study by juxtaposing boxplots of all three testing points (T1, T2, T3). Both in the sphere of content and language, a clearly positive and statistically significant development can be observed. For overall language results, a one-way ANOVA with repeated measures was conducted, showing that the differences in mean scores are statistically significant with  $F(2, 30) = 27.74, p < .001$ ,<sup>97</sup> presenting a large effect size of  $\eta^2 G = .52$ . As T3 did not yield normally distributed results in terms of content,

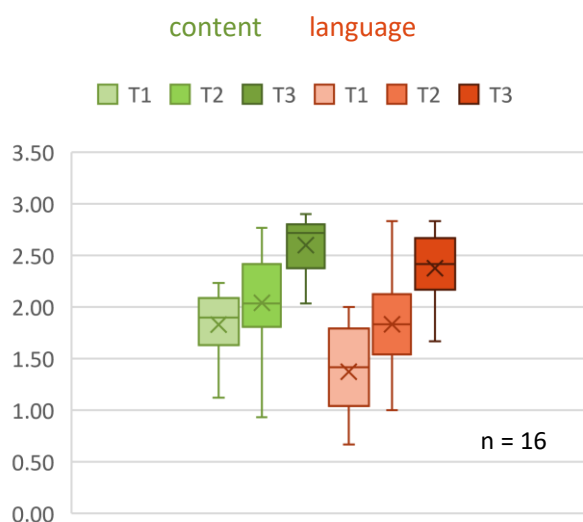


Figure 34. Comparison T1, T2, & T3: boxplots, group A

a Friedman test was performed instead of ANOVA, showing that the changes of average scores are statistically significant ( $\chi^2(2) = 22.63, p < .001, r_w, T2-T1 = .22, r_w, T3-T2 = .59, r_w, T3-T1 = .81$ ).<sup>98</sup>

Figure 34 also shows that at T3, content results of the whole group shifted towards the upper end of the scale. For instance, the minimum value of content outcomes (= 2.03) at T3 almost equals T2's median and mean (= 2.04). T3's content median (= 2.72), on the other hand, is noticeably higher than T1's maximum value (= 2.23) and almost as high as T2's maximum value (= 2.77), signalling that half of

the students performed better at T3 than the best student at T1 and almost as well as the best student at T2. To be more precise, 25% of all students performed better at T3 than the best student at T2 in terms of content. Furthermore, after an initial increase of range and dispersion from T1 ( $R_{T1} = 1.11, SD_{T1} = 0.31$ ) to T2 ( $R_{T2} = 1.84, SD_{T2} = 0.51$ ), these statistical measures decreased again, with  $R_{T3} = 0.87$  and  $SD_{T3} = 0.27$ , indicating that internal differences eventually diminished.

Turning to the linguistic rating, Figure 34 shows that range and dispersion of language results, eventually, decreased too ( $R_{T3} = 1.16, SD_{T3} = 0.36, R_{T2} = 1.83, SD_{T2} = 0.43, R_{T1} = 1.33, SD_{T1} = 0.43$ ). Similar to the content results, but to a less extreme degree, the language results of the whole group improved. In numbers, at T3, 75% of all learners (= 12 students) performed better than the upper quartile (i.e., the four best performing students) at T2. The maximum value remained steady after the second sitting (= 2.83), while the minimum value of T3 (= 1.67) exceeds the lower quartile of T2 (= 1.54), demonstrating that lower-level students could catch up.

<sup>97</sup> The data does not violate the assumption of sphericity as indicated by *Mauchly's Test of Sphericity* with  $\chi^2(2) = 0.76$  and  $p = .145$ , which signifies that "the variances of the differences between conditions are equal" Field (2017, p. 846). Therefore, no correction is needed.

<sup>98</sup> Effect sizes for Friedman tests cannot be directly computed. Thus, a series of Wilcoxon signed-rank tests including effect size calculations have been performed as recommended by Field (2017).

Splitting the students into achievement groups based on their T1 performance, the following developments can be observed (Figure 35). Initially, there were considerable gaps between those three achievement groups in terms of the focal points of this study. Throughout this project, these gaps closed since the different groups all scored similarly at T3. Interestingly, the low-achieving group even managed to outperform their mid-achieving peers. Looking at their different trajectories, the low-achieving group's language development increased sharply throughout the project, whereas their content results first improved marginally from T1 to T2 and only then rose considerably. The mid-achieving group's development, on the other hand, can be described as moderate and gradual. The high-achieving group's development appears to be moderate but consistent, with slightly more gains after the second intervention.

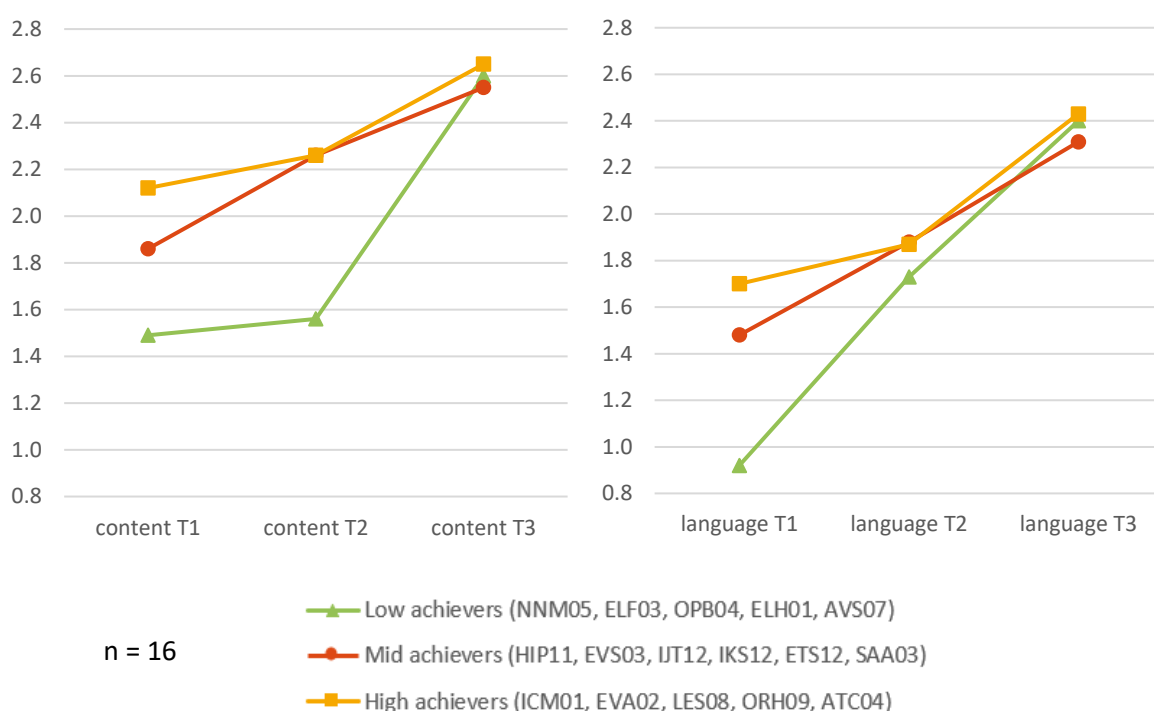


Figure 35. Development of written task results: group A, divided into achievement groups based on T1 performance

As mentioned in section 7.4.2.2, grouping based on T1 performances entails a number of limitations. Therefore, analogous to group B, the students of group A were also split into achievement groups based on their English and history annual grades of the previous year (Figure 36).<sup>99</sup> Starting with content, this graph indicates that all achievement groups performed very similarly on the first written task, but students with higher grades improved more noticeably and more gradually than their peers with lower grades. Additionally, similar to the students with low T1 scores, for students with low grades, the intervention only seemed to take effect in the second

<sup>99</sup> One student (ELH01) did not provide information on this matter and is thus excluded from Figure 36.

half of the project. Turning to language, all three groups improved gradually at rather similar growth rates.

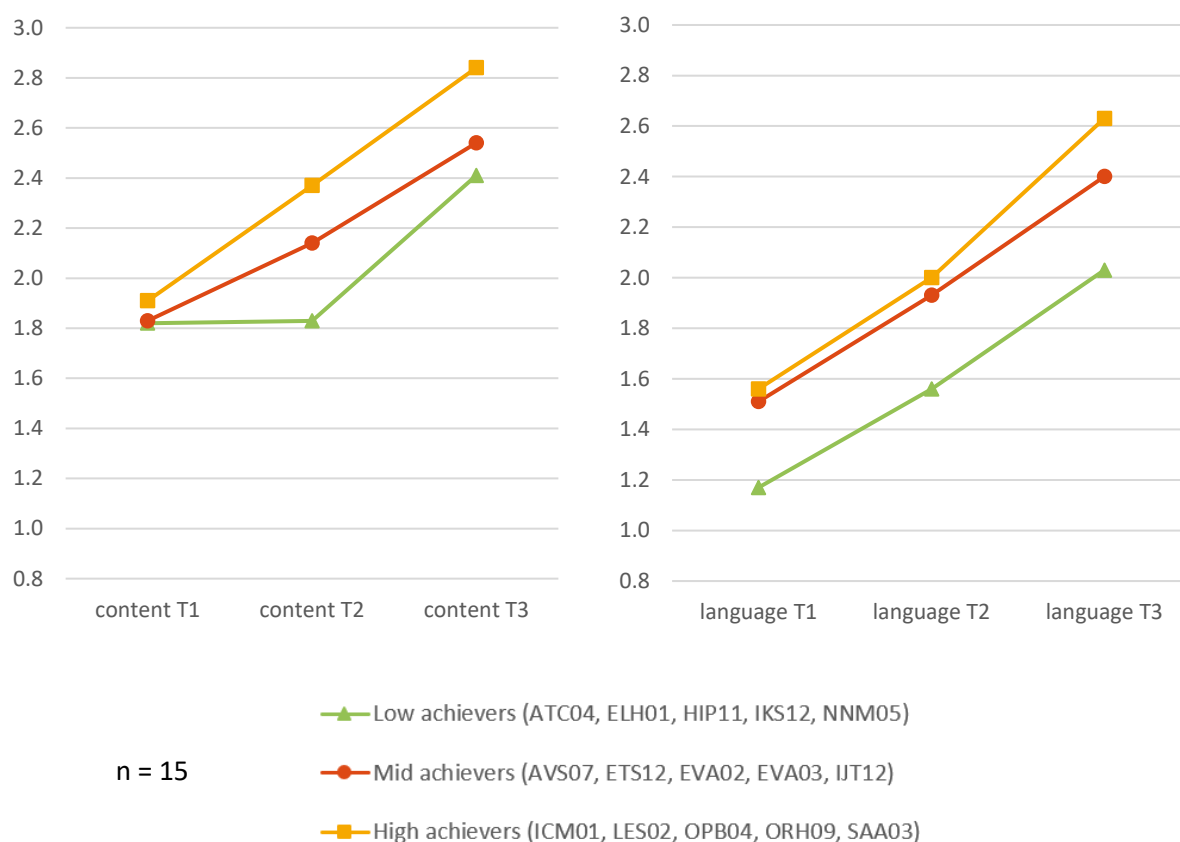


Figure 36. Development of written task results: group A, divided into achievement groups based on previous grades

Now, the difference in defining 'low achievers' and the respective differences in the development could indicate that the intervention was successful in treating the issues identified in the needs analysis, but the intervention is not necessarily more effective for low-achieving students per se. After all, students receive lower grades due to a number of reasons that might or might not be connected to the focus of this study. What needs to be kept in mind, however, is that those students who performed badly prior to the intervention shaped the design to a greater extent. Fortunately, the treatment seems to have worked in these regards, helping these students to improve considerably on the scales used for this study. At the same time, those students who struggled with other aspects that might not be straightforwardly related to focal points of this study and thus received low grades did not benefit more than students with average or high grades. Therefore, it cannot be argued that this intervention closed any proficiency gaps on a broader scale. While it might not be entirely clear who benefitted the most, the data certainly shows that the intervention was indeed effective for the students involved, as their performance of academic language skills and historical competences as measured by the rubrics used for this study significantly improved throughout this project.

### a) History-based rating

Zooming in on the results of the history-focused rating, Figure 37 shows that all areas increased relatively continuously. At T3, the average results of all descriptors are above 2.40, which is more than the highest score at T1 for any descriptor. Furthermore, the gaps between individual descriptors decreased throughout the project, suggesting that certain issues pertaining to some of these areas were resolved for most students. The general tendency of better content results also corresponds to the teacher's perception and her grading of the ongoing school year:

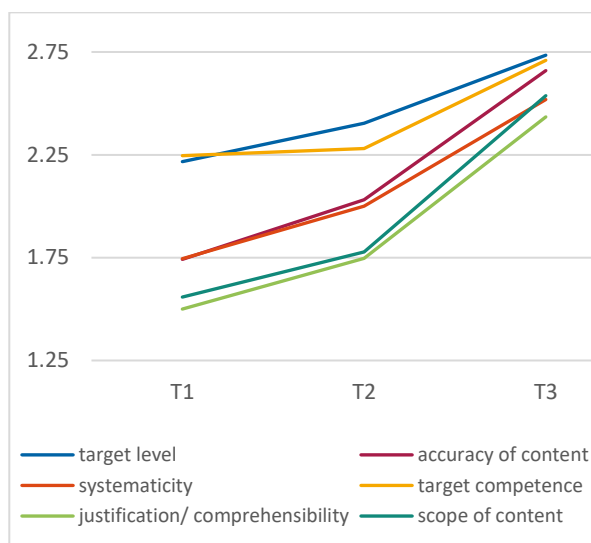


Figure 37. Development of history results, group A

138	English translation	Original quote
T <sub>A</sub>	Yes, this corresponds to my, to the grades relatively well. Apart from ELF03, everybody improved by one grade, actually.	<i>Ja, das, äh, korrespondiert auch mit meinen, mit den Noten relativ gut. Bis auf die ELF03, hat sich eigentlich, haben sich alle um einen Notengrad [...] verbessert.</i>

Thus, the positive changes observed in this study also correspond to real-life success in the subject history.

To check the significance of the results as measured by the written tasks, ANOVA with repeated measure tests were conducted for all descriptors with normally distributed data, while Friedman tests were applied for descriptors with non-normal distribution. Table 16 summarizes the results, showing that all changes were indeed statistically significant.

Table 16. History-based rating: differences between T1, T2, & T3, group A

	Means				Friedman's test			ANOVA		
	T1	T2	T3	T3-T1	$\chi^2$	$F^{100}$	$r_{T3-T1}$	df	$p^*$	$\eta^2_G$
target level	2.22	2.40	2.73	0.52	18.89			2	< .001	
accuracy/ relevance	1.74	2.03	2.66	0.92	16.92		.67	2	< .001	
systematicity	1.74	2.00	2.52	0.77		15.09		2, 30	< .001	.34
target competence	2.25	2.28	2.71	0.46	12.52		.52	2	.002	
justification/ comp.	1.50	1.75	2.43	0.93		28.37		2, 30	< .001	.46
scope of content	1.56	1.78	2.54	0.98		23.84		2, 30	< .001	.48
* $\alpha < .05$										

<sup>100</sup> Calculations on Sphericity: *systematicity*:  $\chi^2(2) = 0.99$  and  $p = .953$ ; *justification*:  $\chi^2(2) = 0.97$  and  $p = .781$ ; *scope*:  $\chi^2(2) = 0.94$  and  $p = .635$ . Therefore, for all three descriptors, sphericity can be assumed.

Table 16 also shows that most progress can be observed in the areas that initially were the weakest, i.e., *justification/comprehensibility* and *scope of content*, followed by *accuracy/relevance of content*. Overall, most learners now managed to justify their answer by connecting their evaluations to the historical context and the content of the picture, as exemplified here:

139	RE: {The people were structured. There was a so called hierarchy were the people on the top of the pyramid had the most power so the pharaoh was allowed to tell the others what they have to do. The people on the top also got more rights than the ones at the bottom.}
Pre-task	
Item 4 OPB04 (B)	EV: {I think it depicts it truthfully, [RE/EA: because as I already said in number three a lot of people moved in cities for different reasons like for example jobs and in this time nobody was responsible, everyone just wanted to make profit (high price for small rooms) & there was not much planning and that are just a few reasons why people ended like the ones in the picture.]}
cycle 3	

In her pre-intervention performance, OPB04 did not evaluate the validity of the source in any way, as she only reported, probably hoping that this report on Ancient Egypt's society would evaluate the authenticity of the image. In her cycle 3 performance, OPB04 clearly articulated her verdict on the validity of the source and argued her assessment by linking the past and the image. Arguably, her chain of cause-and-effect connections is quite dense and could be clearer, but her judgement concerning the validity of the source is convincing. In general, it could be observed that in cycle 3, there were hardly any episodes where students only reported, thereby missing the main point of the task (i.e., the target CDF episode). In their pre-intervention performances, many learners often just reported something they remembered from class, not using it for any other cognitive operations as required by the task. This difference is illustrated in Figure 38:

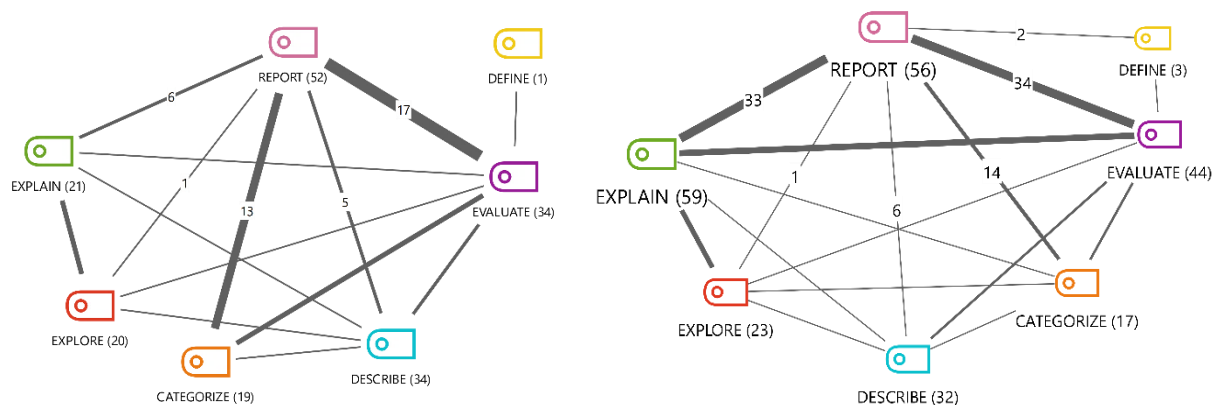


Figure 38. Overlaps of CDF types in the written tasks with a focus on REPORT: T1 (left) vs. T3 (right)

In the pre-intervention texts, REPORT (52) and DESCRIBE (34) are the most common CDF types with relatively few overlaps with EVALUATE or EXPLAIN. While the absolute number of REPORT and DESCRIBE codings in the T3 corpus remained relatively stable (with 56 REPORT and 32 DESCRIBE codings), other target CDF types, like EVALUATE or EXPLAIN, became more frequent and better connected to REPORT and DESCRIBE. Put differently, in cycle 3, many students were not as restricted to CDF types associated with level 1 thinking skills (i.e., reproduction) as was the case pre-intervention, which ultimately improved the learners' history ratings considerably.



Another issue observed in their pre-intervention and first post-intervention performances was presentism, i.e., approaching their source from a present-day bias, thereby neglecting the historical context of the source, which was most obvious in the students' item 2 answers (exploring motives). To address this issue, the cycle 3 materials included an exercise to raise awareness in this regard. Looking at their item 2 answers, it seems that most students made some progress since hardly anybody argued that the main motive of the photographer was to show future generations what life was like in the past. Instead, answers like this were more common:

140	EA/ EO: {Maybe someone might have taken this photo to show how the majority of the families lived back then. Another reason could be to catch richer peoples and the governments attention. Maybe he expected from the government & from richer people to change the situation.}
ELF03 (A)	
Item 2 Cycle 3	
EVA02 (A)	EA/ EO: {To be able to show it in the future or to visualize the effects of industrialisation. The picture could have also been taken because the photographer has a passion for photography. In addition to that do i think that it might have been taken for a newspaper. It also might have been taken to get the government to do something about it.}

After hypothesizing that this photographer took this picture to commemorate the effects of urbanisation, ELF03 considered the photographer's contemporary context and adds two very plausible motives. EVA02 followed a similar strategy and mentioned two rather concrete motives connected to the photographer, speculating about his/her pastime activities or occupation.

#### b) Linguistic rating

Table 17 presents the development of the six different linguistic descriptors used in this study. Similar to the history results, all areas improved throughout the project. Due to the non-normal distribution of the data, Friedman Tests were conducted to test the significance of the changes, showing that all areas except for *composition of CDF types*, which also shows the smallest improvement, were statistically significant. Following up the result for *composition of CDF types* with pairwise comparisons, however, we can see that the differences between T3 and T1 (see below) and between T3 and T2 present a medium effect size ( $r_{w, T3-T2} = .38$ ).

Table 17. Linguistic rating: differences T1, T2, & T3, group A

	Means				Friedman's test			
	T1	T2	T3	T3-T1	$\chi^2$	df	p*	$r_{T3-T1}$
<i>CDFs: choice</i>	1.75	2.06	2.50	0.75	8.67	2	.013	.36
<i>CDFs: composition</i>	1.46	1.50	2.06	0.60	3.60	2	.165	.30
<i>linking/ function</i>	1.25	2.00	2.81	1.56	19.08	2	< 0.001	.72
<i>linking/ form</i>	1.25	1.88	2.50	1.25	16.51	2	< 0.001	.64
<i>hedging</i>	1.00	1.44	1.69	0.69	6.68	2	.035	.38
<i>nominalisation</i>	1.63	2.13	2.69	1.06	13.35	2	.001	.57
* $\alpha < .05$								

Looking at the results more closely, linking seems to be the area showing the most progress. Average ratings for *appropriateness of linking in terms of function* and *form* at least doubled from T1 to T3. In fact, all but two students were rated on the highest level for *linking in terms of function*

at T3, meaning that after the second intervention, almost all students managed to link their ideas appropriately in relation to the CDF type employed. After the two interventions, students seemed to make use of a greater repertoire of linking devices and not just “because” and “so”, which also led to their linking being more precise and fitting to their assumed communicative intention, as can be seen in the following extracts:

141	[EO: I think that the person controlling the workers was someone who either was from the government or has been appointed by them to keep an eye on the workers and to give them instructions. [EV: <b>So</b> , I think that this is connected to the government.][RE: Egyptians founded the irrigation system.]
Pre-task	
ICM01(A) Item 3	
Cycle 3	EV: {I think this source is connected to urbanisation, [DS: <b>since</b> the room on the picture looks overcrowded, [EA/RE: which was <b>a consequence of</b> urbanisation. A lot of people moved from the country side to the city <b>in order to</b> find better jobs or for overall better opportunities. <b>Also</b> , there wasn't much space which also is <b>the reason why</b> it was dirty on the streets and in general.]]}

In the extract from the pre-intervention task, ICM01 used “so” to introduce her explanation why or how the source was connected to centralized governments and/or irrigation. Yet, her hypotheses relating to the people depicted do not really form a tangible basis for her conclusion. Furthermore, her report on irrigation systems seemed quite random and was not linked to the rest. Aside from this “so”, ICM01 only used “because” as connector two times in all five answers (apart from “and” and “or”, which were not counted in the analysis). After cycle 3, ICM01's performance shows a greater range of appropriate linking devices and phrases to express cause and effect. *Linking in terms of form* also shows satisfactory results, with nine students reaching the highest level and only one student remaining on level 1.

Turning to *nominalisation*, significant developments can be observed. After two cycles of intervention, all but four performances were rated at level 3 for their degree of nominal language.

The T3 extracts below show two students with a highly nominal style:

142	[DF: The <b>process of urbanisation</b> describes the <b>movement</b> of people moving from the country side to larger cities, <b>in hope</b> for new/better job opportunities.][EA/RE: This often led to an <b>overpopulation</b> & high rent (in contrast to little wages) in those cities.][DS: The picture shows the living condition of many people [EA/RE: which was a result of <b>urbanisation</b> .]]
Item 3 ORH09	
Item 2 OPB04	EA/EO: {Maybe someone took this picture to show the <b>disadvantages of urbanisation</b> like for example the bad living conditions. Another reason could be to create <b>empathy</b> or <b>awareness</b> .}

In the pre-intervention written task, ORH09 achieved level 2 in terms of *nominalisation* and thus already seemed to have good control over this feature, while OPB04's use of nominalisation was rated as level 1, only using two nominalised forms in total. Looking at ORH09's and OPB04's T3 performances, their writing is formal, dense, and highly nominal, summarizing abstract historical developments.

*Hedging*, on the other hand, started off with low average results and remained the lowest score and the only average value under 2.0 after both interventions. After two cycles, only two students reached level 3 for *hedging*, indicating that this feature of historical discourse appears to be especially difficult for CLIL history learners. Nevertheless, some progress was made, as now no

performance received a zero-level rating (in contrast to seven zero ratings at T1), while level-2 ratings increased from five to seven. Similar to the developments in the students' use of linking devices, learners also seemed to use more hedging expressions and expand their repertoire of different hedging devices. In the pre-intervention tasks, on average, students used three hedging devices throughout their performance, and 50% of all hedging devices were either "maybe" or "I think", i.e., very informal ways of qualifying claims. At T2, students included 3.8 hedging devices on average, with "maybe" and "think" only making up 16% of all devices used. In their final performance, 4.3 hedging devices per student could be observed, with "maybe" and "I think" accounting for 23%. To illustrate, in the pre-intervention task, IJT12 only used "I think" to make her claims less absolute; later she qualifies her evaluations and explorations with epistemic uses of "could" and "would", qualifying adverbial phrases ("does not represent it exactly"), and past modals such as "might have taken".

Throughout the project, the students' *choice of CDF types* became increasingly target-like. At T3, the mode for this descriptor was level 3, with 10 students managing to employ target CDF types for all their CDF episodes and most of their basic CDF types. This is in stark contrast to their T1 and T2 performances, with only two and four level-3 ratings, respectively. Additionally, at T3, there were only two students with level-1 ratings for *CDF choice*. As mentioned before and visible in Figure 38 on page 266, fewer students exclusively opted for reproductive CDF types like REPORT or DESCRIBE when supposed to EVALUATE or EXPLAIN. Instead, students increasingly used CDF types connected to reproductive thinking skills (e.g., REPORT, DESCRIBE) as ways to substantiate their evaluations, explanations, and explorations. In general, EVALUATE, EXPLAIN, and EXPLORE were used more frequently at T3 than T1. These results suggest that most students did not seem to struggle too much with choosing an appropriate CDF type at T3.

As has also been observed in the needs analysis, the results for *choice of CDFs* significantly correlated with results for *target level* and *target competence* ( $\tau_b = .51, p = .024$  and  $\tau_b = .56, p = .013$ ;  $n = 16$ ), i.e., those content descriptors measuring task fulfilment in terms of thinking skill and historical competences. Moreover, results for *choice of CDF types*, together with *composition of CDF types*, were the only descriptors that significantly correlated with overall content results in the cycle 3 data set (*choice*:  $\tau_b = .51, p = .017$ ; *composition*:  $\tau_b = .49, p = .018$ ;  $n = 16$ ), indicating that the use of CDFs is relevant for content performance.

*Composition of CDF types*, however, still appeared to be quite problematic for many learners. The difference between T3 and T1 scores for this descriptor is the smallest of all descriptors and the only difference that is statistically insignificant. After two rounds of intervention, there were still five performances rated at level 1, suggesting that some learners presented a confusing and less purposeful structure of their CDFs with unclear overall communicative intentions. The extract below serves as an example of a complex CDF structure that does not seem to support one main communicative intention:

143	[EV: I think it depicts it truthfully [RE: because People had work very hard, long hours, in bad conditions [EA: there wasn't enough place for all the people as a result of urbanisation.]]] [RE: Hygiene and infrastructure wasn't relevant [EA: they only wanted to make profit so the housing was terrible.]] [EA/RE: In this century many children in young age started working and many people got diseases as a result of low hygiene.]
Item 4	
Cycle 3 NNM05	

In this extract, NNM05 first provided her verdict on the validity of the source, but then she continued with reporting facts about the Industrial Revolution, including cause-and-effect relations without being clear how this would relate to her assessment of the source. Some parts of her report were not even depicted in the source, such as child labour or diseases. Overall, even though she used some connectors, the different basic CDF types were not linked in a way that would support the main point of the task. Instead, the reader is presented with individual bits of information and her evaluations, but they do not seem to be very coherent. It appears that the student wanted to include as many things as possible to demonstrate that she had paid attention during the lessons, but she failed to present it in a way that would be appropriate to the task at hand. Nevertheless, some students did improve in this regard, including six students who managed to reach level 3 (in contrast to zero at T1 and level 1 at T2). An example for such a development can be seen in Extract 141 on page 268 (ICM01, item 3, T1 & T3). In ICM01's T3 performance, the individual basic CDF types support her overall communicative intention, while in her T1 text, neither was it clear what her overall communicative intention was nor how the REPORT-sequence would relate to the other two CDF realizations.

Finally, many of the trends observed in the students' written task performances were also apparent in their reports written during lesson 4 of cycle 3 (part 4). For the most part, these reports were well done. All learners used some of the phrases presented in the language boxes of the materials (e.g., "it seems to me", "additionally", "according to", etc.). Most students employed a formal, nominal style and a range of linking and, to some degree, also hedging devices. The extract below provides an example:

144	<u>Based on my research I would suggest</u> that the government forbids child labour. Children should go to school and study <u>instead of</u> working in factories. <u>In addition, it seems to me</u> that the workers have no security <u>while</u> working. Factory owners should take more care of their workers and treat them better.
ETS12 written report	

ETS12 included many phrases provided by the language boxes (underlined) and constructed a formal and polite texts suitable for the genre required by employing hedging devices, gerund constructions, and linking devices.

Similar to group B's in-class texts, linking and nominalisations were the most prominent features, as these can be found in all reports written by the students of group A. All but four students included some sort of hedging devices. However, like in cycle 2, there are quite a couple of instances where learners used the phrases provided in the boxes but did not incorporate them appropriately into their own writing, as the extract below exemplifies:

145	For me it seems that working in a factory is very hard and dangerous. There is the risk that people especially children or women can hurt themselves. In addition to that based on my research I would, suggest shorter working hours so that the workers do not overwork themselves and are still able to do that work over many years.
NNM05 Written report	

NNM05 seemed to overcompensate, using a great number of the phrases provided (underlined) but which do not necessarily fit the context. For example, “in addition to that based on my research” does not really add to what she reported before but introduces a new move. Interestingly, such constructions could be observed not only in texts written by students that were considered weak but also, to some extent, those that would be considered strong. Overall, however, T<sub>A</sub> was quite satisfied with these texts and assigned a considerable number of extra credits for this task.

### c) Summary and implications

The results of the third set of written performances by group A strongly suggest that the intervention supported the learners in improving in the aspects under investigation, matching findings of other intervention studies such as Breeze and Gerns (2019) or Lo and Jeong (2018). Results for all descriptors improved throughout the project, with 11 out 12 descriptors showing statistically significant changes. In terms of language, scores for *linking (function & form)* and *nominalisation* increased remarkably, while the results for *hedging* and *choice of CDF types* improved moderately. Apart from quantitative gains, learners also seemed to employ a greater variety of linking and hedging devices. Yet, sometimes, the learners did not incorporate these phrases appropriately into their own writing. Scores for *composition of CDF types* only rose slightly (and to a statistically insignificant degree), pointing towards an area that would need more attention in future materials. As mentioned before, this result corresponds to the outcomes of a study by Breeze and Dafouz (2017), who reported that even tertiary students struggled with this aspect. At the same time, this issue also seems to affect content performance. In fact, *choice* and *composition of CDF types* were the only descriptors correlating with overall content results in this data set, tying in with findings by Doiz and Lasagabaster (2021), who demonstrated that complex CDF constructions are a common feature of historical discourse. In the present study, it appears that the ability to perform a CDF type that corresponds to the required historical thinking skill and competence (such as EVALUATE, EXPLAIN, or EXPLORE for deconstruction competence) as well as being able to support this communicative intention with a number of different CDF types relating to the source (DESCRIBE) and the historical context (REPORT) are difficult but necessary moves for source analysis. Therefore, these aspects need to receive more attention in content-and-language-integrative history lessons.

In the content domain, *justification/ comprehensibility*, *scope*, and *accuracy/ relevance of content* saw the highest increases, but all other areas significantly improved as well throughout the project. At the end of the project, all six descriptors reached approximately the same stage, ranging between 2.43 and 2.73 and exceeding the highest T1 score (= *target competence* with 2.25). Therefore, it can be argued that the intervention was successful in tackling the content-related

issues observed in the needs analysis phase, which the teacher also confirmed in her interview. Regarding improvements in the linguistic performance of the students, the teacher was more hesitant, suggesting that her students did not noticeably transfer any of the changes observed in their T3 performance into other situations (see interview with T<sub>A</sub>, subsection 7.4.3.1). This would suggest that the intervention heightened their language awareness and inspired them to use some of the linguistic tools in the context of the study and potentially the subject history on a greater scale, but these changes were not substantial enough at this point to notably affect their general academic language skills. As is to be expected, such an approach would need more time and focus to inspire long-lasting effects that exceed the boundaries of one subject. For comparison, in a study by McCabe and Whittaker (2017) focusing on the development of historical voice, substantial improvements became apparent within a stretch of four years.

Finally, splitting up the outcomes into achievement levels based on the students' T1 performance suggests that the intervention adequately addressed the issues observed in the needs analysis, as the students that initially scored the lowest showed the greatest improvements. This observation matches the outcomes of cycle 2, where students with the lowest T1 scores improved the most. Unlike in group B, the low achievers of group A (as defined by their T1 performance) even outperformed the mid achievers. In fact, by T3, the differences between the achievement groups within group A had disappeared, with all groups reaching approximately the same scores on average. Yet, group A's data would not allow the conclusion that low achievers would especially benefit from this approach, ultimately leading to more equality, since this definition of "low achievers" is very targeted to the focal points of this study. When splitting the group into achievement levels based on their English and history grades, which take into account a variety of achievement markers, the picture somewhat changes. With this way of differentiation, all groups seem to improve at very similar rates, by and large, and no effects of equalisation can be observed. The cycle 3 materials intended to react to previous preliminary insights, which includes catering to diversity by creating optional tasks that are more sensitive to different achievement levels. Since in cycle 3 all achievement groups (as defined by previous grades) improved roughly at the same rate, it could be argued that these measures could have been effective. However, it should be noted that taking diversity into account and catering to different needs was only rudimentarily done in this study and would need further consideration. At the same time, the outcomes of all three cycles point towards the fact that internal differentiation would improve these materials, warranting future research into differentiated instruction for CLIL, as has recently been called for within the CLIL research community (e.g., Calderón-Jurado & Garcia, 2018; Lialikhova, 2019; Pérez Cañado, 2020).

## 8. Discussion

This chapter discusses the results of this study against the backdrop of the literature presented in chapter 2-5. First of all, subchapter 8.1 presents a discussion of the empirical data and the generated design principles, thereby answering the three RQs that structured this research project. Then, subchapter 8.2 revisits the central concept of this dissertation, i.e., the CDF construct, offering a suggestion for a 'history version' of the construct. Finally, subchapter 8.3 discusses methodological insights gained in the course of this PhD project

### 8.1 Answers to the research questions, including design principles

Below, the three RQs will be answered. RQ1, being concerned with learner needs and features of content-and-language-integrative materials, will be answered via presenting design principles deduced from the empirical data of this study and pertinent literature. RQ2 addresses the participants' reactions to the interventions, and RQ3 covers observable effects of the intervention on the learners' academic language and subject-specific skills.

#### 8.1.1 RQ1: learner needs and features of materials

The first research question deals with the type and ideal characteristics of content-and-language-integrative materials and didactic procedures intended to help students improve their expression of cognitive processes when working on historical competences in the context of upper secondary CLIL history education. In other words, the following section presents the design principles (see van den Akker's (1999) definition presented in section 5.1.3, p. 95) developed in this project. These recommended features and procedures are discussed from three different perspectives, namely the learners' and the teacher's perception as well as the researcher's assessment of written student performances, against the backdrop of relevant literature. These are the design principles:

- 1) *Content-and-language-integrative materials should be student-centred, interactive, multi-modal, rich in variation, creative, and tailor-made resources that foster cognitive engagement.*

When asking the learners participating in this study how the intervention should be, both groups listed a number of features which very much align with 'general' principles of good teaching, and more specifically, with principles of good CLIL teaching as outlined in section 2.3.4. For example, the students of group A called for interactive, engaging, creative, tailor-made materials, worksheets, and teaching methods. In case of group A, these characteristics are, for the most part, congruent with their history teacher's lessons. Group B, however, voiced disappointment about their CLIL programme, having experienced hardly any lessons that included the target language in a purposeful and constructive way. As such, they did not wish for the continuation of the status quo and wished for materials that were truly bilingual (i.e., providing space for both languages), student-centred, rich in variation, and multimodal to support comprehension. Many of these

characteristics have been put forward by CLIL researchers and experts on CLIL material design. For example, the importance of student-centredness, interaction, multi-modality, and variation has been emphasized by Ball et al. (2015), Banegas (2017), Mehisto et al. (2009), Meyer (2013), P. Moore and Lorenzo (2015), and Pérez Cañado (2018a) and is also reflected in the principles of the ADiBE Erasmus+ project (ADiBE, 2021).

The two teachers, too, stressed the importance of creating engaging, student-centred, and varied materials in order to counteract the students' lack of interest, involvement, and motivation, which both the teachers and students identified as the biggest challenges in the history classroom. While T<sub>A</sub> did not believe that language played a role in this, T<sub>B2</sub> argued that the reason for affective issues might be language-related, which resonates with results by Somers and Llinares (2018), who have shown that low language proficiency impedes motivation to learn academic content in a CLIL setting (see also Lasagabaster, 2011). T<sub>B2</sub> added that his learners generally present low levels of frustration tolerance and a lack of focus. Therefore, he called for materials that centre on the students' comprehension and help them practise cognitive endurance by making them repeat similar procedures, taking a slightly different perspective each time, which seems to correspond to Meyer et al.'s (2015) call for deeper learning as well as to Tedick and Lyster's (2019) notion of procedural scaffolding for comprehension.

As for interaction and collaboration, the three cycles have shown that the learners appreciate peer work as put forward in their initial interview, but it turned out that there need to be clear instructions concerning the social setting and the distribution of tasks. For instance, the prompt "take notes and discuss with a partner" was too vague as to whether the students should first take notes and then discuss or whether they should take notes together while discussing. In this regard, Tedick and Lyster (2019) underscore the importance of, on the one hand, providing ample space for peer cooperation, and, on the other hand, giving precise instructions when setting up such tasks.

*2) Different scaffolding techniques, including a CDF-based approach, should be incorporated in content-and-language-integrative materials.*

To meet the challenges of bilingual history education, teachers and students expressed the wish for more scaffolding in the lessons. Especially in school B, both teacher and students voiced concerns about feelings of overload, helplessness (see also Otwinowska & Foryś, 2017) and, as mentioned above, little cognitive endurance and would thus welcome more guidance inbuilt in the materials they work with. This is also supported by research into the relationship between CLIL and motivation and affect. While many (early) studies reported a positive CLIL effect (e.g., Doiz et al., 2014; Lasagabaster, 2011; Lasagabaster & Sierra, 2009; Merisuo-Strom, 2007), more recent studies present a more differentiated picture (Mearns et al., 2017; Ohlberger & Wegner, 2017; Otwinowska & Foryś, 2017; Rumlich, 2016; Somers & Llinares, 2018), indicating that a positive link between affective factors and CLIL might depend on the level of support and scaffolding to counteract feelings of overload and frustration.



Yet, highly motivated learners might profit from such an approach as well. Group A, a group that presented relatively high levels of initial motivation, felt that working through content in smaller manageable steps and with more support would help them understand complex input. In particular, historical sources were reported to be difficult to process due to the subject-specific and, at times, antiquated language. Therefore, these learners would welcome a step-by-step approach and the discussion of linguistic features and lexical items. Moreover, the teachers, and to some extent the students too, stressed the importance of guiding learners more explicitly when they actively engage with historical content and verbalize their thinking.

In other words, scaffolding techniques for comprehension (or *input scaffolding*) as well as for production (or *output scaffolding*) should be central features of content-and-language-integrative materials. This has been emphasized repeatedly within the research community (e.g., Coyle et al., 2010; de Graaff et al., 2007; Llinares et al., 2012; Mahan et al., 2018; Mahan, 2020; Meyer, 2013; Meyer et al., 2015; P. Moore & Lorenzo, 2007; Tedick & Lyster, 2019), as it embodies the meaning of the ZPD, a central notion within sociocultural theory, which many assume to be the basic framework for content and language integration (e.g., Banegas, 2013; Coyle et al., 2010; Dalton-Puffer et al., 2010; Donato, 2016; Lialikhova, 2019; Moate, 2010). Following SCT, by guiding learners through the process, cognitive load is decreased, potentially empowering students to reach higher thinking skills than they would without scaffolding (e.g., Donato, 2016; Swain et al., 2015; Tedick & Lyster, 2019). One scaffolding technique for dealing with complex content mentioned by the participants of this study is the use of the L1 (see also Gierlinger, 2015; Méndez García & Pavon Vazquez, 2012; Lasabaster, 2013). This might be a useful complementary strategy to react to and, to some extent, also pre-emptively prevent comprehension issues; yet to support learners in their L2 production, other strategies are required as well.

In this study, the notion of CDFs was chosen to serve as the basis for scaffolding the materials. The students and the teachers reported that learners often struggled with responding appropriately to task and test items, not knowing what was expected when they were asked to ‘explain’ or ‘analyse’, etc. As a result, the learners wished for more guidance in this regard. The teachers, too, felt that making expectations more explicit and guiding learners through the process more overtly in class would help learners cope better in lessons but would ultimately also prepare them more adequately for exams. Looking at the learners’ written performances, it appears that these learners indeed often misread tasks with performative verbs, e.g., by reporting something they remembered by heart rather than doing what was required, matching the results by Dalton-Puffer and Bauer-Marschallinger (2019). Furthermore, Breeze and Dafouz (2017) observed in their analysis of business students’ use of EXPLAIN and DESCRIBE “that the problem is precisely that the student fails to signal either CDF (DESCRIBE or EXPLAIN) explicitly” (p. 88), so that the lecturer involved in this study felt that “[i]n many cases their sentences can't really be understood. You can grasp intuitively what they might have meant if they had expressed themselves clearly” (p. 88). This is also reflected in the results of this study. Often, one might assume that the students’

reporting of historical facts or the description of the visual source might have been intended to evaluate the validity of the source (task item 4) or that describing historical and current circumstances would determine the relevance of a historical source from today's perspective (task item 5), but without linguistically marking these intentions, it remains unclear whether the students were indeed capable of making these connections. One reason for the learners' confusion with regard to CDFs seems to be the teachers' understanding of performative verbs and their functions (or lack thereof), as has been argued by T<sub>B2</sub> as well as by Morton (2020). It seems that different teachers often understand performative verbs differently. The students, too, were aware of this, reporting that in a test they would just write down everything they know "and then let the teacher decide what is right and what is wrong" (IKS12) because "[e]very teacher wants something else" (OPB04).

Yet, the issue does not solely seem to be the 'choice' of appropriate CDFs but how these CDFs are used, combined, and linguistically linked within history as a discipline. For example, when learners were asked to evaluate the validity of a historical source (test item 4), learners often came to a verdict, but their justifications were either non-existent or inadequate, as they failed to link the contents of the source (DESCRIBE) to the historical context (REPORT). Similar issues with constructing subject-specific complex CDF-clusters were also reported in the study by Breeze and Dafouz (2017). Thus, it makes sense to scaffold the process of constructing subject-specific complex CDFs episodes by raising the learners' awareness of what is expected, unfolding the connections between different CDFs and helping students produce such CDF episodes via a step-by-step approach enriched with contextually adequate linguistic tools and explicit instruction. Such (or similar) approaches have also been suggested by Breeze and Dafouz (2017), Breeze and Gerns (2019), Nashaat Sobhy (2018), or Evnitskaya (2019).

Given that not only the use of individual CDFs would need scaffolding, one could also argue that, potentially, an SFL- and genre-based approach might be more viable. However, as has been argued in chapter 3, such approaches do not translate into classroom practice easily, as teachers, especially if they are subject educators, might struggle with the complex linguistic super-structure or would otherwise require intensive training or mentoring (Achugar & Carpenter, 2012; Lo & Jeong, 2018; J. Moore et al., 2018; Schall-Leckrone & Barron, 2018; Schall-Leckrone & McQuillan, 2012). Moreover, genre-based approaches focus extensively on writing and reading, which does not necessarily reflect educational practices in the Austrian context and are thus likely to be rejected by the learners in this and similar contexts. While the two units developed in this study involved more writing than would be typical in traditional Austrian history lessons, they did not focus on complete genres. Nonetheless, for these students, the amount of writing was already borderline acceptable. What is more, working with complete genres is not as flexible as with smaller building blocks, which are also more suitable for oral settings (Dalton-Puffer, 2013; Llinares & Pascual Peña, 2015; Meyer et al., 2015; Meyer & Coyle, 2017). Finally, scaffolding on the basis of CDFs can be connected more straightforwardly to the demands of competency-based

testing as well as to the curricular requirements which are typically formulated via performative verbs. This palpable connection ultimately ensures higher acceptance rates among learners and teachers; a claim that the interview data of this study supports as well. In summary, CDFs offer a flexible base for scaffolding. Yet, this does not negate the possibility of taking a more genre-based approach once learners are more used to writing in history class and focusing on language in content subjects more generally. This could help learners combine these smaller building blocks into larger texts and subject-specific genres effectively, which also resonates with the pluriliteracies approach (Meyer et al., 2015; Meyer & Coyle, 2017).

As for input scaffolding, the textual input used in this study was simplified, elaborated, and pedagogically prepared as suggested by P. Moore and Lorenzo (2007). This included chunking information, guiding attention with visual clues (e.g., highlighting, underlining), adding glossaries, explicit information, and advice on how to go through the material or what to look out for, and directing the learners' engagement with the materials via small, manageable tasks usually just targeting one CDF type. These modifications were met with appreciation by this study's participants.

The scaffolding strategies for comprehension and production incorporated in the materials designed for this study used different techniques such as 'language boxes', 'history tips', visuals, examples, impulse questions, etc. (see subsection 7.2). Yet, scaffolding ingrained in pedagogical material alone does not suffice. Instead, the interviews, alongside the lesson observations, have shown that the teacher needs to actively bring these support measures into being so that the learners would really use and accept them. In other words, the teachers needed to explain why learners should consider these boxes because in the pilot and the first main cycle, students tended to ignore the support measure since they did not understand their relevance for the subject history. Moreover, teachers might want to elaborate some more on the various support measures to ensure that learners indeed understood them. Ideally, teachers should also lay out ways of subject-specific thinking, providing meta-cognitive scaffolding, which, according to Mahan (2020), is mostly absent in CLIL lessons. Scaffolding learner production, in contrast to scaffolding comprehension, has been reported to be rare in CLIL lessons as well (de Graaff et al., 2007; Mahan, 2020; van Kampen et al., 2018). Here, two strategies that could also be observed in the videotaped lessons seem to be important. On the one hand, teachers should model subject-appropriate CDF use (Dalton-Puffer et al., 2018; Doiz & Lasagabaster, 2021) and give learners cues when they are struggling with expressing their cognitive operations. On the other hand, teachers should provide appropriate feedback regarding CDF use, such as clarification requests, recasts, explicit correction, as well as positive reinforcement (see also Tedick & Lyster, 2019).

- 3) *To genuinely integrate content and language, CDFs and form-meaning relations should be put into a subject-specific frame.*

One theme that lies at the heart of this study and was repeatedly brought up in the interviews was how to achieve a genuine integration of content and language learning. As has been argued in

chapter 3, both applied linguists (e.g., Coyle et al., 2010; Lo & Jeong, 2018; Meyer et al., 2015; Ruiz de Zarobe & Cenoz, 2015) as well as subject specialists (e.g., Heimes, 2011; Maset, 2015) call for approaches that better fuse content and language, and CDFs offer a way of organically integrating L2 and subject pedagogies (Cammarata & Cavanagh, 2018; Dalton-Puffer, 2013, 2016; Dalton-Puffer et al., 2018; Dalton-Puffer & Bauer-Marschallinger, 2019; Lorenzo & Dalton-Puffer, 2016; Morton, 2020; Nashaat Sobhy, 2018). Yet, it took a number of revisions and adjustments to successfully reach such genuine integration in classroom practice that both teachers and students would appreciate. For the learners of group A, who were experienced bilingual learners, the initial version of CDF-based scaffolding and the general didactic design “felt too much like English class” (ETS12), rejecting all support measures that focused on lexicogrammatical aspects without making their connection to the history content apparent. Instead, the learners of group A asked for more history-specific scaffolding. Some learners in group B, on the other hand, would prefer more general linguistic support over history-specific language boxes. The difference between these two groups might be that in school B, the learners had not experienced much CLIL prior to the project and also seemed dissatisfied with their EFL lessons and thus would welcome general (academic) language support. Ultimately, the question as to whether to prepare subject-specific or general, academic language support lies in the needs of the students. Support measures only make sense if they support something the learners have not yet mastered. Once the learners are comfortable using the L2 in general educational settings, however, it seems crucial to move increasingly into subject-specific discourse to account for the fact that these lessons are subject lessons as well. What is more, focusing too much on general EFL aspects might not be accepted by the teachers either. In fact, most CLIL teachers do not see themselves as language teachers or even regard any linguistic aspects as their responsibility, including T<sub>A</sub> at the beginning of the project (see also Dalton-Puffer, 2007; Kong et al., 2011; Lo & Jeong, 2018; Morton; Skinnari & Bovellan, 2016; Tan, 2011), which makes it unlikely that they would cater for general language needs. Focusing on CDFs and subject-specific language, however, might be a ‘compromise’ most teachers, such as the two teachers participating in this study, could get behind, given their close connection to curricular requirements and the format of competency-based testing, both of which tend to make use of discourse functions and performative verbs.

Moving on to the integration of content and language within tasks, both groups involved in this study seem to favour tasks that foregrounded meaning rather than form. A similar observation was made by Tedick and Young (2018), namely that students reacted to a focus-on-form intervention with confusion, boredom, little engagement, or strategies to shift the focus back to meaning. Interestingly, the teacher in Tedick and Young’s (2018) study, too, felt that a focus on form in content subjects did not feel right, and similar sentiments could be observed in this study as well. When teaching the task focused on nominalisation, for instance, T<sub>B2</sub> appeared rather insecure, giving different instructions than on the worksheet and then asking me for help. T<sub>A</sub> also felt somewhat uncomfortable, later in the interview stating that this task was “too much”. Similarly, J. Moore et al. (2018) made the observation that such interventions can only work if the

teachers understand the pedagogical purpose of a certain didactic practice, like the reasons for addressing metalanguage (and not just the meaning of linguistic notions embodied in such a practice). Naturally, we do not know whether the teachers' hesitation resulted in the students' rejection of these tasks, whether it was the other way round, or if both teachers and students independently disliked attention to form in history lessons. In any case, in later stages of the project, we tried to always highlight how or why certain linguistic features were relevant for historical discourse by rephrasing some of the language tips, adding further information in this regard, or by encouraging the teacher to discuss these relations in class. Another argument for teaching linguistic features in meaningful subject-specific contexts relates to the students' countless examples of misused 'signal phrases' and cohesive devices in the learners' initial written performances, which also resonates with Donato's (2016) call against pre-teaching and frontloading language in content lessons and T<sub>B2</sub>'s observation concerning the short attention spans of his students.

Similarly, concerning spontaneous focus on form and language more generally in class, the learners of group A were quite irritated (visibly on the videos and reportedly in the interviews) by the teacher asking for "new phrases" rather than for new ideas and mainly providing feedback regarding form rather than meaning. This might have been an overcompensation of the pilot cycle, in which T<sub>A</sub> did not focus on language spontaneously in class, and we agreed afterwards that this aspect should be more pronounced in future cycles. In the final cycle, we decided to focus on the learners' use of language in a more subtle way and more closely connected to the content the students would like to express. This time around, the students were much more satisfied with the balance of language and content, and therefore this point was added to the teacher's version of the materials. The students' rejection of form-focused teaching in content classes suggests that Lyster's (2015, 2007) concept of counterbalance might not be accepted by learners in hard CLIL settings. Considering the contextual differences between typical CBI or immersion programmes, CLIL learners usually use an FL rather than a second language and do not strive for near-native competence (yet). Therefore, including accuracy-based activities might be rejected by learners as well as teachers.

- 4) *To genuinely integrate content and language in the subject history, central subject-specific concepts need to be considered, most notably contextualization and (retro-)perspectivity.*

As has been touched upon above, the analysis of the students' written performances suggests that using the target CDF episode or appropriate linguistic markers does not guarantee a satisfactory answer. Instead, learners need to construct complex configurations of basic CDF types sustaining and fleshing out the overarching communicative intention (see also Breeze & Dafouz, 2017; Dalton-Puffer et al., 2018; Doiz & Lasagabaster, 2021), resulting in "functional stress" as quality criterion for advanced learners of history, according to Lorenzo (2017, p. 36). To be able to do this adequately, the analysis of the students' answers indicate that central subject-specific concepts must be considered, mostly relating to aspects of historical consciousness, such as

contextualization and (retro-)perspectivity; a finding that resonates with Lorenzo's (2017) analysis of cognitive discourse competencies of historical thinking.

When asking learners about subject-specific problems, the first thing they mentioned was a lack of declarative knowledge which is why they struggled with a-chronological approaches and ordering and contextualizing historical events in their timeline. This corresponds to two historical second order concepts as defined by Seixas and Morton (2013), namely *continuity and change* and *historical perspective* (see also section 4.1.2), as well as to one component of van Drie and van Boxtel's (2008) model, namely *contextualization*. In terms of FUER competences, this would mostly affect orientation competence and methodological competence, i.e., the two competences in focus of this study. From an empirical perspective, van Drie and van Boxtel (2008) report that most of the few studies in this area (e.g., Shemilt, 1983; van Boxtel & van Drie, 2004) show that learners often struggle with placing historical events on timelines and contextualizing these events. This has also been observed in the written performances of the learners of this study. For example, when asked to evaluate the authenticity of a source (task item 4), students tended to justify their assessment by describing the contents of the picture without explaining how any of this relates to the historical context of the source, which would be necessary to make a claim about the source's validity. Similarly, the students often ignored the immediate historical context and zeitgeist in their explorations of the reasons why an artist created the visual source the way s/he did (item 2, EXPLORE and EXPLAIN), arguing that the main motive must have been 'to show us how it was'. Although learners claimed that taking over other people's perspectives, past and present, was easy for them, many of their written answers suggest otherwise, as they often judged the past by present-day standards. Here, another often-described phenomenon comes into play, namely presentism (Carretero & van Alphen, 2014), i.e., approaching historical artefacts and content with a present-day bias, such as assuming that people of today's world would have been the target audience when, in fact, most historical sources were produced for people living at that time. Here, history materials should sensitise learners regarding this bias and create awareness of the importance of historical context. In the materials of this study, an activity was developed to initiate this process (i.e., connecting the students' present experience to a past source via an impulse question to heighten their awareness of contemporaneity). In terms of contextualization, materials should provide ample contextualization themselves to help learners orientate and thereby model appropriate presentation of historical content. Moreover, materials need to support learners in expressing spatial and temporal localization via output scaffolding, which might include techniques of backshifting (see Lorenzo, 2017; Lorenzo & Dalton-Puffer, 2016) and explicitly comparing the past and the present.

Another aspect that was mentioned by both groups and the teachers was that learners struggled with engaging with historical sources more generally, which constitutes the core of deconstruction competence of the FUER model as well as other important components of other models of historical thinking/ reasoning (e.g., using *primary source evidence* in Seixas and

Morton's (2013) Big Six model or *use of sources* in van Drie and van Boxtel's (2008) model). Van Drie and van Boxtel (2008) add that most empirical studies show that learners are not very experienced in handling sources and usually do not question their reliability. This can also be found in the data of this study since students often were not very critical when analysing the sources, taking their contents at face value and not linking or contrasting them to the historical context as mentioned above. Ultimately, this rendered the students' contributions superficial and unreflective, which corroborates Lorenzo's (2017) finding that the biggest issue of advanced learners of history seems to be the ability to take a stance and properly and comprehensively justify one's view. In his set of written performances by 16-year-old bilingual learners of history, Lorenzo (2017) observed that learners rarely took a stance and those that did presented "unsubstantiated opinions and without any real analysis or perhaps understanding of its purpose and effect" (p. 37). It appears that learners at this stage do not yet know how they are supposed to take a stance in the subject history and which elements are necessary to make a justified claim. For instance, when assessing the validity of a source, one would need to connect the contents of a source (DESCRIBE) to the historical context (REPORT) and argue to which extent these elements are consistent (EVALUATE and potentially COMPARE); see also the FUER model (Körber et al., 2007) and Bauer-Marschallinger (2016). These expected components need to be communicated to the learners. Considering that these elements can be well described from a CDF perspective, scaffolding for production based on this notion can work in this regard, as has been argued previously. Yet, it seems that merely talking about the importance of these components does not suffice, as some learners indeed mentioned both historical context and contents of the source but still failed to express how they were linked. Here, explicit instruction, perhaps combined with analysis of good examples, as well as more practice would help learners improve in this respect. On a more straightforward linguistic level, the learners of this study argued that the first barrier usually is the archaic language often present in textual sources, and they agreed that the inclusion of a glossary as well as a step-by-step approach working through the source would help them comprehend the text (i.e., input scaffolding).

Finally, according to van Drie and van Boxtel (2008), most empirical studies dealing with historical explanations have shown that producing multi-factorial explanations, adequately considering collective and institutional factors, rather than simplistic, individual explanations often poses a problem for learners (see also Lorenzo, 2017, or van Drie & van Boxtel, 2008). Furthermore, Doiz and Lasagabaster's (2021) results indicate that historians that are less proficient speakers of English rely more on simple and repetitive linking devices, like "because" or "that's why"; a result also found by Llinares and Morton (2017). Such inter-clause linking, of course, expresses causality, but it does not necessarily allow for condensing multiple factors, collectives, or abstract concepts as linking through lexical items would. This, at the same time, requires nominalised phrases (Achugar & Schleppegrell, 2005; Coffin, 2006; de Oliveira, 2010). In the data of this study, most learners, too, realized explanations on the basis of linking devices between clauses and focused on single factors rather than multiple factors combined in abstract

concepts. However, there are some few learners who already presented dense explanations, using asyndetic linking and nominalisations. Yet, it needs to be kept in mind that none of the test items focused on cause-effect relations per se, so more data would be needed in this regard. As for designing CLIL history materials, the literature and, to some extent, the data as well warrant the teaching and practising of subject-specific ways of expressing cause and effect, i.e., history-specific realizations of the CDF EXPLAIN.

- 5) *Upper secondary content-and-language-integrative history materials should address typical features of historical discourse, such as discipline-appropriate linking, hedging, and, to some degree, nominalisation.*

Looking at the learners' initial written performances, the linguistic realization of cognitive operations seems to be particularly in need of improvement since results for *linking*, *hedging* and *CDF composition* were comparatively low. What is more, the results have shown that improving these aspects would not just be for cosmetic reasons. In fact, the results of the linguistic descriptors often showed moderate to strong correlations with content-related outcomes, resonating with Breeze and Gerns' (2019) claim that improvements on the linguistic level indeed promote the demonstration of content knowledge and skills and should thus not be dismissed as superficial.

As mentioned above, the students' written answers presented little variety in terms of linking, usually relying on typical cohesive devices, such as "because" or "so", with little evidence of complex linking or asyndetic structures typical for historical discourse (Achugar & Schleppegrell, 2005; de Oliveira, 2010; Lorenzo, 2017). Moreover, these learners misused typical 'signal words', e.g., by using "because" although they are not expressing cause and effect or "however" without presenting a contrast content-wise. Other students did not use any (or hardly any) cohesive devices. Thus, these answers could not signpost communicative intentions or clearly articulate how certain elements related to each other, i.e., presenting an obscure composition of CDF types (see also Breeze & Dafouz, 2017 or Doiz & Lasagabaster, 2021). Turning to *hedging*, it was observed that many learners did not use any hedging devices at all, delivering their claims as absolute truths, sometimes even emphasising their certainty. This suggests that these learners were not yet aware that historical sources were constructed by other humans with an agenda and that these sources might not, in fact, present objective truths, which ultimately indicates lower degrees of historical consciousness (Körber et al., 2007; Kühberger, 2015). A similar finding was reported by Lorenzo (2017), namely that even advanced learners of history often present their opinions as facts, leaving little room for critical assessment.

Interestingly, in the present study, results for *nominalisation* – despite their importance for constructing multi-factorial explanations, expressing collective agents, and asyndetic structures – did not correlate with content results. At the same time, initial results for *nominalisation* were comparatively good. Thus, nominalisations might not need to take priority when teaching aspects of historical discourse with learners similar to these cohorts. Other studies, however, could



demonstrate the importance of good control of nominalisation for demonstrating subject-related skills and knowledge (e.g., Llinares & Morton, 2010; Lorenzo, 2017; Morton, 2010; Nashaat Sobhy, 2018). Moreover, in analysing expert language including historical discourse, it has been repeatedly shown that nominalisations are central features and have thus been called “gatekeepers” or the “key to unlocking the kinds of vertical discourse through which knowledge is construed in academic subjects” (Morton, 2010, p. 87; see also, for instance, Achugar & Schleppegrell, 2005; de Oliveira, 2010; Lorenzo & Dalton-Puffer, 2016; or Ryshina-Pankova, 2016). For these reasons, nominalisation should not be neglected, but more research in this direction would definitely be warranted. For now, it would make sense to consider first how capable one’s students already are when it comes to adequately producing nominalised phrases and then decide when and to what extent nominalisation should be included in one’s teaching.

As has already been argued above, dealing with linguistic features of subject-specific discourse like linking, hedging, and nominalisations should not be dealt with detached from content. The learners of this study vehemently opposed form-focused tasks and teaching sequences, and the teachers shared this sentiment to some degree. Thus, the connection to the discipline needs to be made explicit to the learners. At the same time, teaching these aspects can also be designed in a more subtle way. For example, one could create tasks that require learners to construct texts in which they have to include certain content points which are, by design, pre-formulated as nominalised phrases. Such tasks would also encourage practising different ways of linking (e.g., “due to”, “lead to”, etc., rather than using “because”; see e.g., task 1b, unit II). More generally, allowing more time for writing and insisting on full sentences from time to time might help learners hone their skills; a view which has also been expressed by the two teachers of this study. Of course, it is neither realistic nor expedient to always ask learners to produce full sentences. It seems that notes or keywords are particularly useful for tasks more prone to spoken interaction, whereas full sentences are more appropriate for cognitively challenging tasks, longer coherent outputs that require more planning and consideration, and for tasks focused on a CDF type the learners need to practise some more.

Nonetheless in oral classroom discourse, teachers should sometimes insist on full sentences too. If necessary, teachers might also want to provide corrective feedback, including awareness-raising remarks regarding the conventions of historical discourse, and should further model suitable structures and elicit self-corrections, as has been suggested by Doiz and Lasagabaster (2021) too. As T<sub>B2</sub> reported, learners tend to provide incomplete answers that are often vague and superficial. Here, it would be important to make them go deeper and express themselves more precisely. Topicalizing language use both reactively and proactively is reported to be rarely done in CLIL classrooms, but, at the same time, it has been argued to be a crucial element in helping learners progress (see Dalton-Puffer et al., 2018; Moate, 2010). Yet, as has been argued before, the interview data of this study suggests that meta-talk and scaffolding for production should be done through the lens of the content subject and its disciplinary language.

6) *Content-and-language-integrative materials should cater for mixed-ability groups to ensure an advantageous development for all learners.*

While the interventions aimed to cater for different learner styles by ensuring a variety of task types, input material, and social formats from the start, different levels of ability were not taken into consideration in early versions. As the study progressed, however, it became apparent that providing for mixed-ability groups was central to the success of the materials. Since the intervention was very student-centred, diverging work paces turned out to be problematic, resulting in fast learners being bored and slower learners feeling overwhelmed and stressed. Both the students and teachers were unhappy about this, and therefore it was decided to include fast-track activities for those that finish early. Given that the scaffolding inherent to our pedagogical design is aimed at supporting low to average achievers proactively, these fast-track activities are designed to be cognitively and/ or linguistically challenging for advanced learners, sometimes with a strong focus on subject-specific linguistic features which are assumed to be new to these learners. Such an approach was appreciated by both stronger and weaker learners. As for its implementation, we adapted our strategy to the needs of the groups as recommended by their teachers. In group B, where learners were more used to autonomous learning and individualised set-ups, we let them work on bigger chunks of tasks, organizing the pair-work phases on their own, and if some time was left at the end, they could also do the additional tasks (source D). Unfortunately, this self-directed organizing of pair-work did not always work as anticipated, with learners doing these interactive tasks rather superficially. Thus, one should plan tasks where learners do not only compare and discuss but where communication and sharing ideas is indispensable. In round three, where learners have less experience with autonomous set-ups and problems with group dynamics, we decided to set shorter but more numerous timeframes with the possibility of doing extra tasks.

Looking at the trajectories of different learner groups throughout the study in terms of the results of the written tasks, it can be observed that low to average ability groups benefitted more substantially than high achievers, resulting in smaller performance gaps. Similar outcomes were also reported in Lo and Jeong (2018), who conducted a genre-based intervention, or Tedick and Young (2018), using a form-focused approach, counteracting the general trend that BE might serve high achievers better than low achievers (see, e.g., Mearns, 2012, Fung & Yip, 2014, and Mewald, 2007). Nonetheless, it needs to be kept in mind that in Lo and Jeong's (2018) study, the post-intervention performance was part of the teaching/ learning cycle and thus was completed with the teacher's help, and Tedick and Young's (2018) study is a qualitative study based on classroom discourse analysis and not statistical analysis. While in the present study the differences of the whole group were tested statistically, a differentiation based on achievement levels did not allow for statistical tests of comparison due to the small sample size either. Moreover, the effect of equalisation only affected group A when ability groups were defined through initial task results and not through more general achievement markers such as school grades. When defining the achievement groups based on previous English and history grades, in

group A, all groups improved similarly, whereas in group B low achievers benefitted most, no matter which definition was used. This implies that, by tendency, the intervention used in this study indeed helped those that struggled with the aspects in focus of this study, which in the case of group B, overlaps with those that are conventionally considered low achievers. For intermediate and high achievers, the intervention turned out to be somewhat less effective.

As a consequence, future materials should provide scaffolding that caters for different learner needs to ensure that all kinds of learners can develop ideally, as has been suggested by Donato (2016) or Lialikhova (2019). Moreover, the interviews with the students showed that the learners were acutely aware of whether they were considered low, mid, or high achievers, which also seemed to affect their motivation. Thus, future materials should make sure that any kind of differentiation is not based on a deficit approach, where strong learners can always accomplish more and are usually considered the ‘donor’ of help and knowledge, whereas low achievers are the ones always receiving help. This would, eventually, lead to a Matthew effect, i.e., which has also been reported in the study by Somers and Llinares (2018). While Lialikhova (2019) fears that heterogeneous grouping runs the risk of silencing or marginalizing weaker learners, deliberate planning could avoid these situations. For example, group work should be planned in a way that requires low achievers, while being adequately supported, to complete a task with outcomes needed by other students too (see CLIL pages in Kilbey et al., 2018, for examples; Tomlinson, 2001). Here, one could first split learners into homogenous groups and adapt the type of task and the degree of scaffolding accordingly (rather than simplifying the input, see also Harmer, 2015; Ur, 2012), and then mix them together again for an information gap activity, for example. For such information gap activities, T<sub>B2</sub> advised to ensure that the tasks require purposeful in-depth collaboration and goal-driven communication. T<sub>A</sub> added that it makes sense to always provide some time for learners to work on their own first to allow weaker or shy students to prepare themselves before having to engage with others so that extrovert and/ or strong students would not dominate peer work phases or classroom discourse too much.

Finally, it should be highlighted again that appropriate consideration of different learner needs and abilities proved crucial for the learners’ approval of the intervention and the learning outcomes measured. In future research, this aspect needs to be considered more thoroughly than was possible in the present study. Indeed, discussions on diversity in CLIL are gaining momentum due to the fact that CLIL is increasingly implemented in unstreamed contexts, i.e., where CLIL is compulsory (see Madrid & Pérez Cañado, 2018; Pérez Cañado, 2020, 2021; Rumlich, 2020). Further research is needed to determine the different needs of diverse learners in a CLIL setting beyond the notion of pace and ability levels and to examine the effect of different types of interventions (CDF-based, genre-based, form-focused, etc.) in mixed-ability CLIL settings. Ultimately, and most importantly, more attention needs to be paid to how one could optimally cater for diverse learners in CLIL settings (see also ADiBE, 2021) and which strategies of differentiation and individualisation prove effective, e.g., homo- and heterogenous groupings,

mastery learning, tiering, peer learning, flipped classrooms (see Smale-Jacobse et al., 2019), or differentiated scaffolding (Donato, 2016; Lialikhova, 2019).

### **8.1.2 RQ2: reactions by learners and teachers**

The second research question addresses the students' response to a CDF-based approach in the context of CLIL, both from their own point of view as well as their teacher's. The most central insights have already been mentioned above, as they have been considered in the development of the design principles. Some general trends and further central insights will be reported below.

#### **8.1.2.1 Perceptions of students**

To begin with, students experienced the approach used in this study as markedly different from their traditional history (CLIL) lessons for a number of reasons. Both groups of learners were initially overwhelmed with the intensity of the lessons, emphasising the amount of writing, the dense structure, the increased use of the FL, and that focusing on subject-specific language felt unfamiliar at first. However, all groups seemed to agree that they were more active than usual and also felt that these lessons helped them learn and progress, especially in terms of history but also in terms of language. Looking at the learners' language produced in class and on their worksheets, the scaffolding seems to have positively affected their language use, which was increasingly visible in the students' use of phrases provided by the language boxes, including more instances of hedging and nominalisation. Hasenberger (2018), who is also currently conducting a CDF-based DBR study, has, so far, observed something similar since his students perceived a learning benefit both concerning content knowledge and language skills. Likewise, Nashaat Sobhy's (2018) students in her CDF-based intervention study reported that the focus on defining helped them retain and express content knowledge more effectively.

For some, particularly the students of group B, the language boxes, glossaries, and other support measures were considered to be instrumental in solving the tasks. This compares with the students' evaluations of a genre-based intervention study in Hong Kong by Lo and Jeong (2018), whose learners felt that working with model texts and explicit instruction regarding connectives helped them write a better text. For group B, the linguistic support did not, pre-eminently, help them improve their output per se but instead was perceived to be a necessary tool to work through the tasks in the first place. Not having experienced much CLIL prior to the study, some students of group B had trouble coping with the increased use of English combined with perceived higher levels of difficulty, reflecting sentiments found in a survey of CLIL beginners by Broca (2016); but see also Pérez Cañado (2012), Smit and Finker (2018), or Pladevall-Ballester (2015). More proficient students, especially in the first round of the intervention (group A), however, felt patronized by these boxes, limiting their expression. Such feelings were also voiced in the intervention study by Tedick and Young (2018), working with a form-focused counter-balance approach. In their study, high-intermediate and advanced learners "experienced frustration and confusion early on [...] perhaps due to their well-developed 'intuitive' grammatical systems that

did not align with the overly simplified explanations presented by teachers and researchers” (Tedick & Young, 2018, p. 314), which could also apply to the more advanced students in this study. On closer examination of the interview data and the lesson transcripts, however, it seems that the dissatisfaction of some students with the linguistic support might have to do with the classroom implementation of these measures rather than the materials per se. As mentioned before, in the first cycle, the teacher focused on the students’ use of the phrases provided in these boxes, asking for “new phrases” instead of further ideas in terms of content and regularly providing corrective feedback. In this regard, Döring (2020) recommends “spark[ing] students’ motivation by creating a safe environment in which they are praised for successful communication and task achievement, rather than penalized for language mistakes and partial inability to display content knowledge in an FL” (p. 9), based on a small-scale survey of Austrian CLIL students in upper secondary vocational education. In other words, to ensure that learners do not reject focusing on (subject-specific) language within the context of hard CLIL content lessons, it makes sense to prioritize meaning over form and to consider linguistic aspects more subtly. In group B, where the teacher only rarely addressed these aspects in classroom talk, the students were more content with the linguistic support measures. Moreover, trying to include meta-talk with more care and linking corrective feedback to meaning rather than form was greatly appreciated by group A in the third cycle. Interestingly, by this time, the students were considering the provided phrases and explicit instructions of their own accord. During group and individual work sessions, some students even asked their peers and their teacher actively for feedback on their subject-specific use of language, also in terms of form. This suggests that there was also a learning curve regarding how to deal with meta-talk and how to balance language and subject matter in classroom discourse, not only for teachers but for learners too. This reflects recommendations by experienced DBR researchers, such as Eijkelhof (2017), stressing that any innovation takes time to get accustomed to and therefore advising against control groups in early stages of design research. Edelson (2006) agrees and adds that questions about effectiveness should not be asked “too early or too often” (p. 104), accounting for the exploratory and innovative nature of DBR.

What all learners seemed to appreciate were the type of tasks and the learner-centredness they embodied, which corresponds to the results reported by Oxbrow (2018), whose respondents expressed appreciation of task-based CLIL approaches. The students of the present study also valued, for the most part, the clarity of the prompts, highlighting colour-coding strategies and clear layout. In terms of educational value of the tasks, most learners appeared to be satisfied. However, for some, especially the more gifted learners, some tasks of unit I were perceived to be too repetitive. Moreover, some argued that it felt too similar to EFL instruction. Instead, these students called for more subject-specific approaches and tasks that present novel content (both conceptual and linguistic) to keep it interesting and appropriate for history lessons. As for the overall structure of the materials, most learners seemed content not only with the overall outline but also with the sequence and general variety of activities in the materials of both units. However,

more proficient learners perceived some tasks of unit I to be too predetermined by the sequence of tasks and the small steps. Interestingly, however, when talking about this issue, these learners asked for more flexibility when “summarizing” sources, indicating that they, in fact, were not yet sufficiently familiar with the methodological script and purpose of source analysis, which would require more than just summarizing information; a finding present in the data of Lorenzo’s (2017) study as well.

Another issue that came up in all three retrospective student interviews is linked to the wish for more differentiated instruction. Especially after the first cycle, students complained about diverging paces, annoying fast learners and overwhelming weaker learners. Reacting to this criticism, we planned to individualise the learning process in cycle 2 by making part 4 (source D) optional and letting learners work through the absolutism worksheet in their own pace. While this neutralized diverging work paces, it did not ensure that the cooperative tasks were done in sufficient depth and detail, as pointed out by the students of group B. In cycle 3, we again adjusted our strategy in this regard to the needs of the group, introducing fast-track activities. This was generally met with appreciation by low, mid, and high performers alike. As a consequence, the learners of group A were much more positive about the approach of this study after cycle 3. As such, it appears that individualised and differentiated instruction played a central role in ensuring student approval of the materials developed and working bilingually in general. This seems to support the viewpoint of Somers and Llinares (2018) that better and more differentiated scaffolding could help increase the motivation of learners with lower proficiency and motivation towards CLIL. As has been argued before, the role of differentiated and individualised instructions should be explored in further research to really exploit CLIL’s full potential, which resonates strongly with the ADiBE (2021) project (see also Pérez Cañado, submitted).

To summarize, initially, the learners in this study were rather sceptical about the intervention, unlike the lower-secondary pupils in Lo and Jeong’s (2018) intervention study, who were quite enthusiastic throughout. However, listening to the learners’ criticism and incorporating their feedback has made the students participating in this study considerably more positive towards this approach. In group B, who expressed very negative attitudes concerning CLIL in their needs analysis interview, students appreciated CLIL more after the intervention than before, even though they would prefer a less intense version of it. In group A, where students experienced two rounds of intervention, one could observe a considerable attitudinal shift, with learners valuing the approach much more after the final cycle. In other intervention studies like Ohlberger and Wegner (2017) or Connolly (2019), no such effects could be reported, which, the authors argue, could be explained by the novelty of the approach as well as feelings of overload. While these factors also seem to apply in this study, especially in the context of group B, the perceived benefit of the approach appears to overrule these sentiments. In contrast to the present study, Ohlberger and Wegner (2017) and Connolly (2019) mentioned above only included a one-time intervention. In the study at hand, the design team could gradually improve the intervention, considering the

learners' voices throughout. As has been put forward by Coyle (2013), Döring (2020), and Filice (2021) in the context of CLIL, and by Cook-Sather (2006, 2020), Flutter and Rudduck (2004), Groundwater-Smith and Mockler (2016), or Mitra (2018) more generally, students are capable of – and also appreciate – contributing to the improvement of their own education. Given that they are the target audience, it is only logical to consider their wishes and ideas when creating new didactic materials and tools. Moreover, from the perspective of DBR, taking the participants' voices into account is assumed to be a key element for the success of the design (Dijkstra et al., 2017; Lo & Jeong, 2018; McKenney & Reeves, 2012). The results of this study seem to corroborate this viewpoint since it appears that incorporating participants' voices in the planning of didactic innovation supports participant approval and ultimately improves the perceived and, as section 8.1.3 will later indicate, actual learning outcomes.

### **8.1.2.2 Perceptions of teachers**

The perceptions of the teachers regarding their students' reaction to explicit teaching of CDFs in history lessons partly corresponds to the students' account. Both teachers acknowledged the labour intensity of the unit and confirmed that such an approach, including the increased amount of writing and focus on language, was completely new to these learners. The learners also felt that the materials were rather challenging for their students, stretching the limits of low achievers while not boring high achievers. For these reasons, T<sub>B2</sub> often included recaps in German, which has been observed as a frequent strategy to ease cognitive load in CLIL (Meyerhöffer & Dreesmann, 2019). Another strategy of this teacher was to relate the content to their present experience, e.g., by linking historical content to current politics or soccer. In terms of engagement levels, T<sub>B2</sub> felt that low-performing students might have participated less, while stronger students were more engaged than in traditional lessons; an assessment that seems accurate when looking at the submitted worksheets, even though the interviewees and respondents of the feedback survey reported that they participated more than usual. In context A, the teacher agrees with the learners that everybody was more engaged than in traditional lessons, which was one of the main benefits of the approach from T<sub>A</sub>'s point of view, especially since lack of engagement was one of her main concerns at the outset of the project.

Regarding learning outcomes, T<sub>B2</sub> felt that the students mostly benefitted in terms of linguistic expression while content learning did not suffer. T<sub>A</sub>, being an English teacher, could not observe any tangible long-term language gains, arguing that the intervention was too short for such an effect. In terms of subject matter, T<sub>A</sub> agreed with her students that the intervention helped them improve their history skills, which was also visible in better grades than usual. For T<sub>A</sub>, this mainly has to do with the learner-centredness and competency-orientation of the tasks as well as the clear structure of the materials, the prompts, and the small steps, helping the learners approach sources more systematically. She also highlighted the usefulness of the linguistic support measures, especially those that dealt with performative verbs. T<sub>B2</sub> agreed and, generally, emphasized the value of the linguistic support strategies too.

As for their own experience, both teachers reported that they enjoyed teaching these units and would continue to use these materials and the approach as such, yet in less dense way. As with the students, however, it appeared that it took some time for the teachers to get used to the consideration of subject-specific and academic language in their content classes, as they were struggling with finding an appropriate balance between form and meaning in classroom discourse. This somewhat also resonates with the views of the teacher involved in Lo and Jeong's (2018) intervention study, who argued that teachers "may need to be 'psychologically' and 'practically' prepared" (p. 43) before implementing a new language-focused approach to content teaching. To ease this process, the teacher's versions of the materials developed in this study were annotated in detail to help teachers use the materials, which T<sub>B2</sub> appreciated explicitly. By the second implementation in the context of group A, the teacher already appeared much more attuned to the new approach, which she confirmed in her interview, describing the experience as pleasant and easy.

### 8.1.3 RQ3: the effect on learner language and content learning

The third research question concerns the effect of CDF-oriented teaching on the learners' demonstration of historical competences and academic language skills as observed in the learners' written performances, comparing pre- and post-intervention results. For this purpose, two assessment rubrics were designed; one based on the CDF construct and one based on the Austrian guidelines for history testing and the FUER model. In subsection 8.1.3.1, results of super-categories (*content* and *language*) are briefly presented, before zooming in on content-related insights (8.1.3.2). Then, subsection 8.1.3.3 discusses linguistic outcomes in more detail. Finally, subsection 8.1.3.4 reviews the trajectories of different achievement groups.

#### 8.1.3.1 Overall results: content and language

Looking at whole-group results of the super-categories (*overall content* and *language*), group A, having experienced two interventions, presents relatively steady and statistically significant increases in both areas. Especially in terms of language, the growth rates are striking in this group, showing a plus of 73% throughout the project (history: +42%). In terms of content, the learners of group A started on a relatively good level already, before increasing slightly after the first round of intervention and subsequently making a considerable leap in the final cycle. In the case of group B, who only participated in one cycle and who also had had only little CLIL experience prior to the project, only language results increased significantly (+ 23 %), whereas content-related results virtually stayed the same. This ties in with trends reported in a systematic literature review by Graham et al. (2018), who argue that factors like previous bilingual experience might considerably affect content-related outcomes. Essentially, group B did not only undergo a CDF-based intervention but also a CLIL intervention more generally. As such, group B's zero content benefit corresponds to the overall observation that CLIL, by and large, does not affect content learning, while language learning is often reported to benefit from the approach (e.g., Dallinger et al., 2016;



Dalton-Puffer, 2008; Pérez Cañado, 2012; San Isidro & Lasagabaster, 2019). Group A, having had more experience with BE generally and with this intervention more specifically, could improve in both areas, but again, linguistic gains were more substantial, at least as measured in these task performances.

As for the distribution of outcomes, in both groups, standard deviation and absolute range decreased from the pre-intervention results to the final post-intervention results, suggesting that, overall, differences between the learners diminished somewhat. Another general interesting observation concerns the connection of *overall content* and *overall language* results. In all data sets, *content* and *language* results correlated moderately to strongly, supporting the assumption of a conceptual link between the two dimensions, as has been argued in Bauer-Marschallinger (2016) and Dalton-Puffer and Bauer-Marschallinger (2019) in the case of history but also, more generally, in Morton (2020) or Nashaat Sobhy (2018).

### **8.1.3.2 History outcomes**

Zooming in on the history results, it becomes apparent that although both groups started from a similar point of departure, the different areas under investigation, i.e., the six history-related descriptors, developed quite differently in these two groups. Initially, both groups presented the lowest scores in terms of *justification/ comprehensibility* as well as *scope of content*, while *target level* and *target competence* were the strongest areas in both datasets. So, relatively speaking, before the intervention, all learners participating in this study tended to perform the target competence on an appropriate level of historical thinking (reproduction vs. reorganisation/transfer vs. reflection/ problem-solving), but they failed to provide details and justifications for their views and claims. In the case of group A, these two weak points presented the sharpest growth throughout the project, ranking similarly to the results of the other descriptors at the end of the project. Overall, the final content results of group A are extremely satisfactory, with average values ranging between 2.4 and 2.8 (max = 3.0), suggesting that in potential follow-up studies, a more nuanced scale might be needed. In group B, however, *justification/ comprehensibility* only improved marginally, and *scope of content* even decreased somewhat. Here, the post-intervention performances were still rather superficial, lacking justifications and details for claims (or other central content elements). This corresponds to Lorenzo's (2016) as well as J. Moore et al.'s (2018) observation that even advanced learners of history struggle with articulating justified claims and comprehensible analyses. Considering that these issues were more salient in the items that asked the learners to evaluate, these insights also relate to the findings by SFL-based studies looking into history learners' appraisals. For example, Morton and Llinares (2018) reported that those learners with better language skills could better express appraisals and develop appropriate voice, while lower-rated learners struggled. Similarly, group A, who started with a similar overall rating but improved substantially throughout the study, also increased their *justification/ comprehensibility* and *scope* scores accordingly, unlike group B. Yet, especially in context B, where learners only participated in this project over the

course of a month, a large effect in this regard would be unrealistic, considering that in McCabe and Whittaker's (2017) longitudinal analysis of voice development, considerable growth of these skills became apparent over a stretch of four years.

Interestingly, *justification/ comprehensibility* turned out to be the descriptor that correlated the most with other descriptors, both in terms of content and language, highlighting how notionally important this aspect is within historical discourse. Here, especially *composition of CDF types* and *systematicity* stand out, both conceptually and statistically. As has been argued above, for a complete evaluation of a source, learners often need to relate the contents of the source to the historical context (DESCRIBE, REPORT, CATEGORIZE). Considering that various elements need to be assembled logically and systematically for this purpose, it is not surprising that learners struggled in this regard. Looking at the trajectories of the two groups in this dimension, these challenges are mirrored in the results for both descriptors. As for *systematicity*, the learners of group A started out quite strong ( $M = 2.25$ ), potentially due to the reason that their answers were often short and thus not difficult to systematize. Then, after the first round, learners indeed tried to justify their claims, yet they struggled with presenting the various elements, i.e., basic CDFs, coherently, which manifested in a decrease of *systematicity* scores on average ( $M = 2.00$ ). After the final intervention and some more practice in this regard, the *systematicity* scores went up again ( $M = 2.41$ ), but it nonetheless remained the lowest in the set. In group B, *systematicity* scores increased similarly to the scores for *justification/comprehensibility*, yet these changes were very subtle. Other intervention studies, in contrast, suggest that systematicity and organisation are areas very susceptible to improvement, such as Lo and Jeong (2018) or Schleppegrell et al. (2008). These SFL-based studies, however, are focused on specific genres; therefore, it is not surprising that structure and organization improved more tangibly in these studies.

Another issue observed in the initial performances relates to the notion of presentism, i.e., the learners' failure in viewing and presenting the source in its historical context (Carretero & van Alphen, 2014; Seixas, 2017), which was also visible in Lorenzo's (2017) data. This relates most straightforwardly to the descriptor *accuracy/ relevance of content*. In the case of group A, this area increased moderately after cycle 1 and considerably after cycle 3, analogously to the overall development. In group B, there is only very slight growth, also matching the overall trend of this group. Taking a more qualitative perspective, issues with presentism were most obvious in answers to task item 2, i.e., speculating about motives and reasons why a person created the historical source in question. Especially after the third cycle, where we actively included an activity targeting such thought processes, the learners' answers to this item were considerably more sensitive towards historical contexts.

Finally, *target competence* and *thinking level* remained the best-rated descriptors throughout the set, suggesting that at upper secondary level, learners are already quite capable of demonstrating methodology competence and orientation competence at the appropriate level of historical thinking, corresponding to the results presented by Bauer-Marschallinger (2016) and Dalton-

Puffer and Bauer-Marschallinger (2019). Nonetheless, there were a number of answers in which learners only reproduced content covered in class or obvious in the source rather than relating different aspects or performing critical analysis or reflection, affecting their *target level* results. In these cases, learners tended to REPORT memorized chunks from class or DESCRIBE what was depicted in the visual source instead of evaluating, explaining, exploring, or categorizing; a result also found by Bauer-Marschallinger (2016) and Dalton-Puffer and Bauer-Marschallinger (2019). As such, this descriptor often correlated with the descriptor *choice of CDF type*. Similar to *target level*, the results of *choice of CDF type* were (among) the highest in the pre-intervention results, which then improved moderately, remaining in the higher ranks of the post-intervention results.

### 8.1.3.3 Language outcomes

Moving on to language outcomes, the most notable result is that in both groups all individual descriptors increased eventually, except for *composition of CDF types* in group B, which stayed at exactly the same level. As with content results, both groups started at rather similar levels and with similar weak areas, those being *hedging* and *linking*. Linking, especially in terms of *function*, is the descriptor presenting the most substantial increase. In the case of group A, average ratings more than doubled throughout the project, ending up being the highest score in the final data set. Ratings for *linking in terms of form* also rose considerably in this group. In group B, this area presented the largest growth with a plus of 55% for *linking/ form* and 44% for *linking/ function*. Unlike in group A, their final scores for *linking* are not the highest of the set (which was *choice of CDF type*). Taking a qualitative perspective, in the initial performances, some learners did not link their ideas at all, or only loosely, whereas at later points in time, they used suitable and more precise cohesive devices much more frequently. Moreover, they used a greater range of linking devices after the interventions, both in the post-intervention tasks and the in-class tasks. Similar changes were also observed in the intervention studies by Breeze and Gerns (2019) or Lo and Jeong (2018). Concerning the quality of answers, Breeze and Dafouz (2017) found that low-proficiency learners often failed to adequately and explicitly link ideas, especially connections between sources, theoretical concepts, and their own thoughts, which would be necessary to comprehend their reasoning; a phenomenon also clearly present in the data of this study.

As for *hedging*, scores in this domain started on a very low level and remained the weakest area after the interventions in both groups. Nonetheless, some headway could be made. In the context of group A, scores increased by 80% and in group B, where only one cycle took place, a growth of 22% could be observed. Initially, most learners did not hedge their claims at all, often presenting their views and interpretations as absolute truths, which is a phenomenon also observed in the study by Lorenzo (2017). Alternatively, these learners used the same rather simple phrases again and again. After the interventions, subtle changes could be observed in the data of this study. Learners used some of the phrases and strategies that were introduced in the material, increasing their range of expressing caution, both in the post-intervention tasks and their in-class outputs. However, the students sometimes struggled with appropriately integrating these strategies into

their own writing, suggesting that hedging is an area that would need more time and practice, especially in view of its importance for historical discourse (Coffin, 2006, Martin & White, 2005, Lorenzo, 2017). Again, this is not surprising, considering that in a longitudinal study by McCabe and Whittaker (2017), only the higher-rated group hedged their claims sometimes, showing “a slight tendency towards an opening-up of the dialogic space” (p. 118) after a period of four years. In Lorenzo’s (2017) analysis of advanced learners’ historical literacy skills too, problems in this area were observed across the board.

*Nominalisation* scores, on the other hand, started relatively high and remained strong throughout the project in both groups. In group A, all but four final post-intervention performances were rated at level 3 in terms of nominalisation. In group B, learners used a more nominal style compared to their initial performance, albeit with a higher frequency of mistakes and errors. Interestingly, outputs produced in class often presented a very nominal style, indicating that the learners did consider the language support, even though in the interview, they rejected the idea of having any input or focus on nominalisation. In other intervention studies, such as Achugar and Carpenter (2012), Breeze and Gerns (2019), or Lo and Jeong (2018), learners also managed to use a dense, more nominal style after the intervention. Even in contexts with little focus on language, history learners seem to adopt an increasingly nominal and complex style, according to a longitudinal study by Whittaker et al. (2011). Collectively, these results suggest that history learners often implicitly take on a more academic, nominal style when confronted with historical discourse, but some additional attention might push learners to improve faster. As has been stressed, for instance, by Nashaat Sobhy (2018), Morton (2010), Achugar and Schleppegrell (2005), or Lorenzo (2017), nominalisations are central for credible and appropriate academic output. In connection to CDFs, Nashaat Sobhy (2018), in her study operationalizing the CDF DEFINE, observed that weaker students’ definitions usually lacked nominalised structures, and thus she recommends including this aspect in CDF-based teaching.

Finally, the use of CDFs, being the prime focus of the intervention, has been moderately affected in context A and marginally in context B. As has been mentioned before, *choice of CDF types* started relatively strong in each group and, following slight to moderate increases, remained an area with comparatively high scores. What is quite striking is that in group A, after two rounds of intervention, the mode for *choice of CDF type* was 3, indicating target-like use of individual CDFs across the board. *Composition of CDF types*, on the other hand, was identified as a weak domain in the pre-intervention phase and could improve to some extent in group A, while group B did not change at all in this domain. At the same time, being able to compose logical compositions of CDF episodes has been found to be a crucial skill in historical discourse (Doiz & Lasagabaster, 2021) and other disciplines, posing a challenge even for tertiary students (Breeze & Dafouz, 2017). From a qualitative point of view, the data of the study at hand has shown that adequate choice and logical composition of CDF types are crucial elements in successfully demonstrating history skills. More specifically, they seem to be a pre-condition for meeting the requirements of competency-

based tasks, including demonstrating thinking skills at the target level and the ability to justify one's claims. These links are underpinned by statistical analyses showing that these two CDF-related descriptors tended to correlate with overall content results, which suggests that adequate CDF use is linked to adequate subject performance. Similarly, Breeze and Gerns (2019) have observed that the learners in their study produced more complete answers from a content perspective after the intervention, indicating as well that a focus on expressing cognitive processes indeed helps learners demonstrate their content knowledge successfully.

#### **8.1.3.4 Achievement groups**

Splitting the groups into achievement groups based on their pre-intervention task performance shows that learners with the lowest initial scores benefitted most from the intervention in both groups. In the case of group A, those with the lowest scores in the first task even outperformed the average group at the end of the project. In fact, in group A, all three achievement groups (as defined by T1 performance) performed similarly well in the final task, suggesting that the gap has closed, at least in the areas covered by the rubrics and in focus of this study. Yet, when grouping learners into achievement groups based on their English and history grades of the previous year, these trajectories change considerably. Using this criterion to differentiate developments, low, mid, and high achievers fared equally well, improving their scores at similar rates, which also means that the gap was maintained. In other words, these results indicate that the intervention was successfully targeted at the experimental group, effectively helping them to work on areas identified as weak within the focus of this study, while those that were already quite skilled in this regard did not change much. At the same time, when using a broader and also more conventional definition of achievement groups (i.e., previous grades), one could not really argue that learners with lower achievement levels more generally would benefit most. Interestingly, in group B, both the learners with the lowest grades and the lowest T1 results improved the most. The results of mid and high achievers, on the other hand, only increased marginally or even decreased in some domains, depending on the definition of the achievement groups. In any case, it needs to be kept in mind that the sample sizes of these achievement-based groupings are small, and thus these results need to be treated with caution. Nonetheless, it appears that a language-based intervention supports those that struggle with expressing subject-specific skills adequately; a result which has also been reported by Lo and Jeong (2018) using a genre-based approach or Tedick and Young (2018) employing a form-focused approach. In contrast, when measuring the impact of EMI instruction without taking linguistic support measures, like in the study by Fung and Yip (2014), low-achieving learners tend to fare worse than their mid- and high-achieving peers (see also Mearns, 2012, or Mewald, 2007).

## **8.2 Conceptual discussion: the CDF construct revisited**

This subchapter discusses what the results of this study mean for the central concept of this dissertation, i.e., the CDF construct. To begin with, this study has shown that the CDF construct is

a useful tool to operationalize the integration of content and language learning in the CLIL classroom, confirming similar claims by Nashaat Sobhy (2018), Morton (2020), or Breeze and Gerns (2019). Since CDFs present a finer granularity than genres (Dalton-Puffer, 2013), they could be easily worked into the 'flow' of the subject-specific content, adequately reflecting the learning objectives (as set by the curriculum and chosen by the design team). As such, CDFs are indeed a flexible and versatile concept to "conciliate teaching content and language" (Nashaat Sobhy, 2018, p. 110) on smaller scales, like individual discourse functions, but also on a larger scale via their combination into more extensive episodes or even complete subject-specific texts, depending on the learners' needs and contextual factors (Meyer & Coyle, 2017; Nashaat Sobhy, 2018).

What is more, due to their resemblance to performative verbs, CDFs have face validity with both teachers and students. Clarifying the target communicative intention of performative verbs and practicing their enactment is seen as purposeful by the teachers and, for the most part, students too. Thus, the construct lives up to its promise of combining the perspectives of content and language pedagogies without "being experienced as transgressive or even meaningless" (Dalton-Puffer & Bauer-Marschallinger, 2019, p. 33) but "as being 'theirs', serving the interests of their subject and actually being part of what doing history [...] involves" (Dalton-Puffer, 2013, p. 242). It seems that in hard CLIL settings like in Austria, learners especially, but teachers to some extent too, reject an overt focus on language in content classes (along with extensive writing). CDFs, however, are 'close enough' to the contents and learning goals of the subject and therefore work well as a 'compromise' to pay attention to subject-specific and academic linguistic concerns in a way that is accepted by all participants. At the same time, working with CDFs makes the linguistic demands of content learning objectives visible and more relevant for subject teachers, who often do not consider language as part of their responsibility (see also Evnitskaya, 2019; Morton, 2020). On a smaller scale, CDFs can also help make task requirements more tangible, eventually resulting in more target-like and precise student performances (Breeze & Dafouz, 2017; Morton, 2020; Nashaat Sobhy, 2018).

Taking a research-methodological perspective, I agree with Lorenzo (2017, p. 40) that Dalton-Puffer's (2013, 2016) construct is a "comprehensive but also manageable taxonomy to conduct research on language across the disciplines" (p. 40, see also Breeze & Dafouz, 2017). Yet, as Doiz and Lasagabaster (2021) have rightfully pointed out, "the characterizations of some CDFs need to be slightly adapted to capture the particularities of the discipline" (p. 67) to be able to use it reliably as an analytical tool. Considering that the original construct set out to present flexible categories whose realizations can be adapted to the requirements of the discipline (Dalton-Puffer, 2013), further specification for the different disciplines, ideally via the collaboration between subject specialists and linguists, seems to be well within the meaning of the construct. Taking the perspective of the history educator and applied linguist in personal union, the following presents an attempt at specifying Dalton-Puffer's (2013, 2016) CDF construct for the subject history. Table 18 on page 298 presents an overview of the results. These specifications are based on the

qualitative analysis of the data of this study, but this conceptual analysis also considers previous (theoretical) elaborations on historical literacy as outlined in chapter 3. Moreover, it pays special attention to CDF- and history-specific analyses, e.g., by Lorenzo (2017), who analysed essays of 10<sup>th</sup>-grade bilingual learners, and by Doiz and Lasagabaster (2021), who analysed EMI university lectures. As such, these specifications largely rest on learner language, albeit at fairly advanced levels. To solidify the history-version of the CDF construct as presented in Table 18, it might therefore make sense to revisit these suggestions on the basis of L1 language productions and/or a greater variety of L2 output in future research projects.

Apart from specifications concerning the communicative intentions underlying the CDF types, this conceptual analysis attempts to specify typical linguistic features of these CDFs in the subject history. Obviously, resting on learner language and some of the literature in the field as presented in chapter 3, this list of features is not meant to be exhaustive.

Starting with CATEGORIZE, this CDF has been split into COMPARE and CLASSIFY. As has been argued by Evnitskaya and Dalton-Puffer (2020), comparing is much more common in historical discourse than classifying or categorizing. The data of this study, i.e., the written learner performances, the discussions with teachers, and lesson observations, suggest the same. Learners and teachers frequently compare past and present, different sources, or contents of a source and their historical context (which often co-occurs with DESCRIBE, REPORT, or EVALUATE). Interestingly, even though CATEGORIZE was not defined as a target CDF episode and only once as a target basic type, it appeared surprisingly frequently in the learners' written productions, and their use usually made sense in the composition of the learner. In the lessons, too, comparisons were often included, and the teachers explicitly mentioned in the design sessions and interviews that comparing would be a central function in need of more attention. Nonetheless, both classifying and categorizing are important acts in academic discourse (see, e.g., Beacco, 2010; Kidd, 1996; Mohan, 1986; Trimble, 1985) and can appear in history too, which is why COMPARE did not replace CATEGORIZE or CLASSIFY. Instead, CATEGORIZE serves as super-category to the related functions COMPARE and CLASSIFY because it is epistemologically broader and more flexible than CLASSIFY. In other words, CATEGORIZE includes both elements of comparing and grouping rather than only working with systematic classes (see Ellin, 2004, and subsection 3.4.4.2). While such a relation has been implied in past publications (Dalton-Puffer & Bauer-Marschallinger, 2019; Evnitskaya & Dalton-Puffer, 2020), it has not been included in the construct itself. In terms of linguistic features, being able to express contrast and similarities is a necessary skill. Moreover, nominalisations and abstractions help systematize these aspects into groupings or classes.

Table 18. Revision of Dalton-Puffer's CDF construct for the subject history

<b>CDF</b>	<b>general communicative intention</b> (Dalton-Puffer, 2013)	<b>specifications for the subject history</b>	<b>linguistic features relevant for this CDF in the subject history</b>
CATEGORIZE: COMPARE & CLASSIFY	I tell you how we can cut up the world according to certain ideas	<i>I tell you about similarities and differences (COMPARE)</i>	nominalisations and abstractions expressing contrast and similarities
		<i>... and how we can cut up the world according to certain ideas (CLASSIFY)</i>	
DEFINE	I tell you about the extension of this object of specialist knowledge	<i>...in the context of its time</i>	nominalisations, abstractions
DESCRIBE	I tell you details of what I can see (also metaphorically)	<i>I tell you details of what I can perceive on the basis of historical sources and materials</i>	referring to parts of sources and describing historical entities using adequate vocabulary
EVALUATE	I tell you what my position is vis a vis X	<i>... (e.g., the validity or historical significance of a source, an argument, an opinion, etc.) and I provide you with historically valid justifications for this view</i>	differentiating between fact and opinion, using different hedging devices; justifying views by, e.g., comparing past and present, contents and style of a source and their historical context, or different sources (corroboration)
EXPLAIN	I give you reasons for and tell you cause/s of X	<i>I give you reasons for and tell you about the causes or motives of X</i>	causal linking; multifactorial & asyndetic linking; abstractions and nominalisations
EXPLORE	I tell you something that is potential (i.e., non-factual)	<i>I tell you something that is counter-factual (= sth. that could have been) or speculative (= sth. that might have been)</i>	hedging; expressing counter-factuality, hypotheticality, and speculation (modality, conditionals)
REPORT: NARRATE & SUMMARIZE	I tell you sth. external to our immediate context on which I have a legitimate knowledge claim	<i>NARRATE: I tell you sth. external to our immediate context, i.e., not observable in the sources/ materials at hand, on which I have a legitimate knowledge claim</i>	linking, backshifting, navigating textual time, nominalisations
		<i>SUMMARIZE: I give you a condensed version (= key points) of what I have been working on recently</i>	linking, organisation, nominalisations, abstractions



Moving on to DEFINE, the underlying communicative intention now specifically covers the temporal dimension, as suggested by Doiz and Lasagabaster (2021). In history, specific concepts such as *manufacture* or *bourgeois* and general terms like *power* or *politics* often carry a present or everyday meaning and a history-specific or temporarily-bound one (see also *terminology competence*, Schöner, 2007, and section 4.2.4). While this meaning has not been excluded from the original construct, it makes sense to specifically include it in this version of the construct given the importance of temporality in historical definitions. Linguistically, definitions – as in other disciplines – require abstractions and nominalisations (Lorenzo, 2017; Nashaat Sobhy, 2018).

Working through my own data and looking at other researchers' mappings of DESCRIBE, it appears that DESCRIBE needs further specifications for the subject history. For Doiz and Lasagabaster (2021), the needed revision for this CDF type as well as for REPORT would be adding a “temporal component as teachers frequently relate past and present time events and situations” (p. 68). There are three reasons why I do not find this suggestion expedient. First of all, such a temporal component was not excluded in the original construct, at least not to my understanding. Moreover, relating past and present would rather fall into the CDF type CATEGORIZE/COMPARE. Finally, their suggestion concerns both DESCRIBE and REPORT; two CDF types that, in the subject history, tend to be difficult to tell apart based on the original construct.

Looking at the analysis of Doiz and Lasagabaster (2021), I would not always agree where they assigned DESCRIBE and REPORT, and in discussions with other researchers working with the construct, the difference between these two does not seem to be defined precisely enough. So rather than expanding both, it would make sense to differentiate them more clearly from the perspective of the discipline. Taking the FUER competence model and other models of history skills that centre on source analysis and the construction of historical narratives, I suggest the following: DESCRIBE is concerned with details and perceivable properties of historical sources and/ or materials at hand, trying to grasp and adequately display what we perceive, most often in order to deconstruct and examine historical sources (see *deconstruction competence*, Schreiber, 2007b, and section 4.2.2). To do this successfully, learners need a rich vocabulary (see also Lorenzo, 2017) and phrases to refer to parts of sources. REPORT, on the other hand, covers all acts of recounting anything “external to our immediate context” (Dalton-Puffer, 2013, p. 234), i.e., not observable in the sources or materials at hand. Such acts are linked to re-construction competence or narrative competence, as the elements previously extracted from historical materials should now be comprehensively and reasonably combined into one historical narrative (see *reconstruction competence*, Schreiber, 2007b, and section 4.2.2). To clarify with an example, DESCRIBE would apply when describing the physical properties of a historical source or a historical event, person, or item as depicted in historical materials, whereas REPORT would apply if someone outlined a historical event, person, or item from memory or after having worked through a number of sources or texts, reporting what they find important about the historical entity in question. To allow for more precision, this type of REPORT has been termed NARRATE, reflecting the

notion of the historical narrative prominent in history didactics. As has already been pointed out in previous studies (Bauer-Marschallinger, 2016; Dalton-Puffer & Bauer-Marschallinger, 2019; Lorenzo, 2017), narrating is a central skill in historical discourse, which usually involves synthesizing information from different sources. This is not trivial, both cognitively and linguistically, and may require the ability to combine several basic CDFs into one longer episode or even into a full genre such as historical report or narration. This, in turn, rests on skilful linking and logical organization of information as well as the ability to backshift, navigating historical and textual time (see also Lorenzo, 2017). Narrating, however, might not be the only type of REPORT in history lessons. More generally, learners (or historians) are often asked to give a condensed version of what they have been working on recently, which is now a sub-type labelled as SUMMARIZE. To SUMMARIZE, one needs good linking skills and the ability to abstract and use nominalisations.

Concerning EVALUATE, I suggest two additions. First of all, the importance of justifying one's view should be reflected in the construct. While this aspect is by no means excluded from the general construct, it appears that history educators and all notions of historical consciousness, thinking, and reasoning stress the significance of justifying one's claims in a historically valid way (Körber et al., 2007; Rüsen, 2004; van Drie & van Boxtel, 2008). At the same time, learners struggle with this greatly (see also Lorenzo, 2017). Thus, it makes sense to include this aspect explicitly in the 'history version' of the CDF construct. Secondly, in line with Doiz and Lasagabaster (2021), it should be clarified that taking a stance concerning someone else's evaluation, which is a prominent element in historical discourse, especially in higher grades and tertiary education, counts as well. Thus, the element to be evaluated, i.e., *X*, has been exemplified to make clear that evaluating *reported* evaluations also counts as EVALUATE. However, and this seems to be in disagreement with the analyses by Doiz and Lasagabaster (2021), if other people's views are only reproduced without taking a stance concerning their view, then this is not necessarily a sign of historical maturity. In fact, just presenting someone's view shows that these learners have not yet grasped the *particularity principle*, i.e., that all historical narratives and judgements are constructed, contextually-bound, and biased in one way or another (Körber et al., 2007). Similar lines of argumentation can also be found in Rüsen's (1983, 2004) description of historical consciousness or Van Drie and van Boxtel's (2008) understanding of historical argumentation, all stressing the importance of carefully gauging historical interpretations and claims. Thus, when referring to other people's views without tangibly providing one's own take, I would argue that this constitutes an instance of NARRATE (REPORT) rather than EVALUATE. Linguistically, historically valid stance-taking requires the ability to differentiate between fact and opinion, using different hedging devices to show that other interpretations might also be valid and using language for justification (appropriate linking, expressing reasons, abstractions). Moreover, this might further entail comparing past and present (COMPARE), contents and style of a source and historical context (DESCRIBE, REPORT, COMPARE), or comparing different sources (*corroboration*, see Lorenzo, 2017,

but also Reisman, 2012; DESCRIBE & COMPARE). As such, sophisticated evaluations are prone to functional stress and are likely to function as CDF episodes.

Turning to EXPLAIN, only slight changes have been made. The underlying communicative intention now explicitly includes *reasons*, *causes*, and *motives*, combining the original wording (Dalton-Puffer, 2013) and the wording found in Dalton-Puffer and Bauer-Marschallinger (2019). Language-wise, learners need to be able to express causal and multifactorial relationships, which, in turn, require asyndetic linking and nominalisations (see Achugar & Schleppegrell, 2005; Lorenzo, 2017). In the data of this study, instances of EXPLAIN were often interlaced with REPORT (e.g., when relating established facts), EVALUATE (e.g., when assessing the significance of a historical development), and EXPLORE (e.g., when speculating about potential reasons or hypothesizing about future developments).

EXPLORE has also been slightly rephrased to highlight that thinking about hypotheticals and contingencies does not only concern the future, which is rather rare in historical discourse (Doiz & Lasagabaster, 2021; Lorenzo, 2017), but also counter-factual thought-experiments (“what-ifs”) and speculations about the past, i.e., things we can never legitimately determine from today’s perspective, such as feelings of historical figures. As such, EXPLORE tended to co-occur with REPORT, EXPLAIN, and EVALUATE in the data of this study. Concerning linguistic demands, learners need to be able to express counter-factuality, hypotheticality, and speculation, for all of which students need to make use of modality and conditionals (see also Bauer-Marschallinger, 2016; Coffin, 2006; Dalton-Puffer & Bauer-Marschallinger, 2019; Lorenzo, 2017).

Finally, the data of this study confirms once more that CDFs do not only run on one level but several, as has been put forward by Breeze and Dafouz (2017), Dalton-Puffer et al. (2018), Dalton-Puffer (2016), and Doiz and Lasagabaster (2021). In this study, learners very often sustained a larger and often more complex communicative intention (*episode*) with various smaller CDF elements (*basic*).

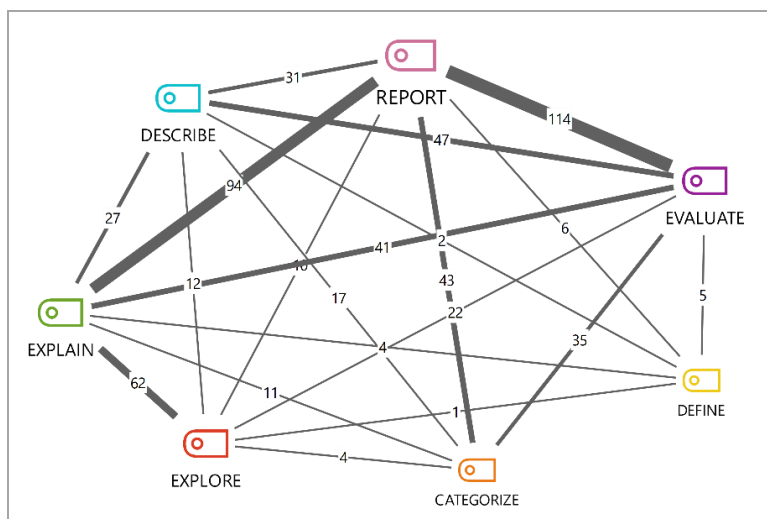


Figure 39. Code co-occurrence model of all CDF codings of all written task performances

The code co-occurrence model above (Figure 39), which includes all written data of this study, illustrates this. Additionally, the qualitative analysis of such CDF clusters has shown that successful episodes are characterized by a logical and clearly linked assembly of the different parts.

### 8.3 Methodological insights

The most central insight regarding research methodology is that design-based research indeed presents a viable approach to “effectively bridge the chasm between research and practice in formal education” (T. Anderson & Shattuck, 2012, p. 16, see also Barab & Squire, 2004; Euler, 2014; The Design-Based Research Collective, 2003; Wang & Hannafin, 2005). In the context of this study, practice and research were coupled on various levels: First of all, the materials produced in this study were targeted at alleviating a practice-related problem, i.e., the lack of appropriate CLIL materials that integrate content and language learning, which was not only reported by local practitioners, i.e., my former colleagues and the teachers participating in this study, but also in the literature (see section 2.3.3 and Banegas, 2014; Hahn, 2019; Massler, 2012; Meyer et al., 2015; Morton, 2013; Pérez Cañado, 2018). Secondly, the materials developed made use of theoretical notions assumed to link content and language learning and also considered the practice-related concerns voiced by both teachers and learners. In fact, the teachers co-designed the materials to make sure they are context-sensitive and ecologically valid. Finally, designing the materials did not only result in a collection of didactic resources but also in a continuous development of theoretical insights, producing the design principles outlined in subchapter 8.1, following van den Akker’s (1999) understanding of design principles, and the conceptual deliberations presented in subchapter 8.2. As suggested by McKenney and Reeves (2014), this study did not primarily focus on “what works” but “how we can make something work and why” (p. 143).

In order to develop materials that worked in terms of learning outcomes and participants’ approval, three cycles in total were necessary, and two rounds per group led to better results than one round per group, both in terms of learning outcomes and approval rates. In other words, as suggested in the DBR literature (Euler, 2014; McKenney & Reeves, 2012; van den Akker et al., 2006a; Wang & Hannafin, 2005), one-time interventions are not enough to adequately solve educational issues, even if the intervention is research-based and the participants’ perspectives are considered. One reason for needing at least two cycles per group relates to the type of input learners can share at different points in the research process. In the needs analysis interview, the learners only voiced general ideas and features which they would like to see in content-and-language-integrative materials, but they could not really express specific suggestions. Before this research project, they simply could not imagine how one could integrate content and language learning save for glossaries. After the intervention, however, their inputs and feedback were much more nuanced and focused on the main point of this study. Put differently, in early interviews, participants contributed general ideas and expressed tendencies, whereas at later stages, their input helped fine-tune the materials and conceptual insights. Therefore, I recommend at least two research cycles per group.

While learners might not have been able to provide detailed and targeted input prior to the intervention, their contributions were crucial for the success of the design. Listening to the learners’ voices turned out to be a central element in the fine-tuning process of the design,

improving the materials, as has been suggested by Döring (2020), Coyle (2013), or Filice (2021) in the context of CLIL (see also Cook-Sather, 2006, 2020; Flutter & Rudduck, 2004; Groundwater-Smith & Mockler, 2016; Mitra, 2018). Moreover, the findings of this study suggest that listening to the students and developing materials accordingly did not only result in more positive learner attitudes towards the materials but also in perceived learning benefits and, potentially, actual learning gains, at least as measured by the written tasks used in this study.

As has been demonstrated in this thesis, the development of the materials, the design principles, and the theoretical considerations has been a gradual process, mostly relying on methods of qualitative research (and the quantification of these findings illustrating the development and the outcomes). Moreover, teachers and students needed time to get used to the new approach. For these reasons, having control groups would not have been appropriate in this phase of the design, agreeing with Eijkelhof (2017) as well as Euler (2014) or McKenney and Reeve's (2012) conceptualization of alpha- and beta-testing in DBR. Once the design is ready to move into the so-called gamma phase, in which the design developed in this study would be now, randomized field trials with control groups, allowing statistical generalization and determination of large-scale effectiveness, would be suitable next steps (Euler, 2014; McKenney & Reeves, 2012).

Another central methodological insight concerns the typical setting of DBR, namely real-life classrooms. While it makes sense to develop educational designs in naturally occurring test beds to ensure "responsively grounded" outcomes "so that we can explore, rather than mute, the complex realities of teaching and learning, and respond accordingly" (McKenney & Reeves, 2012, p. 15), conducting research in schools entails a number of challenges that are not always easy to anticipate. Tight and changing schedules, learner absences, unforeseen events, incidents, or tests in other subjects are only some examples of what may and indeed has affected the research presented here. Of course, to some degree, having experienced these factors makes the educational design more ecologically valid, but, at the same time, it also impacts the quality of the data. For these reasons, it is vital to communicate clearly with the participating teachers, as recommended by McKenney et al. (2006) or Kelly (2006). Based on my experience, I would also add that it makes sense to stay in touch with the class teacher, as sometimes they know about events or changes in schedules that the content teachers do not. Furthermore, it makes sense to plan sufficient time between cycles to have a time buffer in case of unforeseen changes and to ensure enough time for the researcher to work through the data and for the design team to implement these insights accordingly. Additionally, implementation phases should not be too close to school holidays to avoid having an unexpected break in the middle of the intervention in case of unanticipated changes in the schedule. Moreover, with holidays approaching, students might struggle with focus and motivation. Obviously, sometimes it is not possible to fully consider these aspects, but it would indeed reduce the pressure while collecting data in schools.

Finally, as has been partly discussed in subchapter 8.2, the CDF- based as well as the history-based rating rubric allowed a reliable assessment of the data, considering that intra-rating correlation

was found to be strong (see subsection 5.5.3.3). Yet, it took two rounds of analysis in the pilot phase plus a complete first round of analysis of the data of the main study to create comprehensive and appropriate rubrics and a coding guide to ensure a reliable final rating of the data of the main study. As a result, these rubrics and the coding guide are rather detailed and are, therefore, likely to be impractical for teachers in their assessment of learners. At the same time, having three-level rubrics might not be nuanced enough, especially for learners with good results in these aspects (cycle 3). Moreover, these analytical tools have been designed for the purpose of this study and thus focus on specific aspects of historical literacy and only a selection of history skills. For these reasons, these tools would likely need adaption if used in a different context with learners at different levels, both when used for teaching and research purposes. Nonetheless, they could present helpful starting points for further research but also for CLIL assessment.

## 9. Conclusion

This final chapter concludes this thesis by recapitulating central themes of this PhD project. First, a summary of the study at hand is provided, reviewing various stages of this thesis (9.1). Then, key findings and implications are presented (9.2). Finally, in subchapter 9.3, the significance but also the limitations of the research presented are discussed, concluding with recommendations for future work.

### 9.1 Summary

This study set out to develop content-and-language-integrative CLIL history materials for the upper secondary level while also furthering our understanding of the interconnectedness of content and language learning. As outlined in chapter 2, CLIL research has, so far, mostly focused on linguistic outcomes and to a noticeably lesser extent on its impact on content learning. For a long time, most research in the field was conducted by applied linguists, while subject-specialists were not as involved in CLIL research since they tended to be sceptical about the surplus value of the approach for their subject. This dichotomous view has recently started to dissolve, reconceptualizing CLIL as an approach in which language and content learning objectives overlap rather than co-exist. Although the research community appears to agree on the importance of theorizing the integration of content and language learning, few attempts have been made to operationalize such findings for classroom use. At the same time, research into the beliefs and practices of CLIL teachers has shown that CLIL practitioners struggle with understanding their own role and responsibility when it comes to linguistic aspects. To complicate matters further, there is a paucity of adequate CLIL teaching resources generally but especially in terms of materials that integrate content and language learning. Turning to the learners' perspective, the majority of CLIL research has dealt with the learners' motivation towards the foreign language, neglecting affective factors and beliefs relating to the content subject and the integration of content and language learning. Additionally, CLIL research has recently started to pay more attention to heterogeneity in CLIL settings, considering that in many countries CLIL has been introduced into mainstream educational contexts.

With this in mind, chapter 3 explored different theoretical approaches that organically integrate content and language integration and subsequently discussed their usefulness for classroom practice in the context of history education. Starting with sociocultural theory (SCT), although this theory presents a helpful framework to understand learning in CLIL, it does not straightforwardly translate into the didactics of specific subjects. Systemic functional linguistics (SFL), on the other hand, provides a more tangible as well as a potentially subject-specific approach to integrating content and language learning and teaching. Here, based on the notion of genre, concrete pedagogical approaches have been developed, such as the teaching/learning cycle or Reading2Learn. However, genre-based pedagogy centres on written language, which often plays a minor role in European CLIL programmes, and also requires a fair amount of linguistic

sophistication on the teacher's part, which only seems to work if teachers attend comprehensive in- or pre-service teacher trainings or participate in mentoring programmes focused on SFL-based teaching. As a result, it is unlikely that teachers, especially those without linguistic training, take up such an approach also because many of them do not feel that teaching language beyond subject-specific vocabulary would be part of their responsibility. In contrast, Cognitive Discourse Functions (CDFs) present a high degree of face validity due to their conceptual link to the performative verbs used in many curricula as well as assessment. Moreover, CDFs present a finer granularity than whole genres and can thus be flexibly integrated into the flow of content teaching, which in Austrian CLIL classes mostly happens in the oral mode. In a number of studies, Dalton-Puffer's (2013) construct of CDFs has been empirically validated, yet only a few studies have been conducted that operationalize the construct for classroom use. In this regard, Dalton-Puffer (2013) stressed the importance of transdisciplinary work and the inclusion of researchers with a background in the content subject.

With the author of this dissertation having a degree in both English language and history teaching, the role of applied linguist and subject educationalist has been conjoined into one person in this study. As such, central notions, themes, and theories of history education were reviewed in chapter 4 to situate this study not only within the sphere of educational linguistics but also history didactics, allowing for a transdisciplinary approach. More specifically, this chapter also outlined how history learning is conceptualized in the Austrian context by examining the competency model underlying Austrian secondary curricula, namely the FUER model, as well as the policies regarding history education in the context of this research project.

Considering that this study aimed at both advancing the theoretical underpinnings of content and language integration and exploring ways of operationalizing these insights for pedagogical practice, taking a transdisciplinary approach does not only entail integrating the perspectives of educational linguistics and history education. Instead, it makes sense that researchers and practitioners join forces to create solutions that are research-based and ecologically valid. Design-based research (DBR) appears to be a methodological approach that provides a fitting framework for such a study. In chapter 5, this approach is first introduced and discussed on a general level before outlining the specifics of the study at hand, including the context of the study, the research design, the research questions, and methods of data collection and analysis. Zooming in on the methodology of this study, two classes of upper secondary history education (group A and B), taught by two different teachers (teacher A and B2) at two vocational schools in Austria (school A and B) participated in this research project. Typical for DBR, this study was organized in research cycles. Such a cycle started off with a thorough analysis of the participants' needs on the basis of semi-structured individual teacher interviews, focus group interviews with students, and written learner performances, which were elicited using a prompt targeting the demonstration of the two history competences in focus of this study as well as a range of CDFs. In light of the insights gained, the respective teacher and I collaboratively created pedagogical interventions incorporating CDF



theory. These design sessions were audiotaped to document our thought processes and the development of the materials. The practitioner then put these materials into practice in their own classrooms. Subsequently, the process and product were formatively evaluated, using retrospective interviews with the teacher and a group of students as well as written tasks once more. The first cycle took place in context A before moving into context B for the second cycle. Here, too, the needs of the group were identified using the same procedures as in the first cycle. These findings and the results of the first research cycle informed the revision of the intervention, which was then re-implemented by teacher B2, which was followed by another evaluation process. In accordance with the results gained thus far, our approach was refined and applied to a new topic to be implemented in the first school (A) again. Finally, this last cycle was evaluated, again using teacher and student interviews as well as written tasks.

Concerning data analysis, the transcripts of the interviews and the design sessions were analysed via qualitative content analysis according to Kuckartz (2016) with the help of MaxQDA, using a deductive-inductive approach for the interviews and a mostly inductive approach for the design sessions. For the analysis of the written tasks, two rubrics were developed; one based on the CDF construct and one based on the Austrian guidelines for competency-based history testing, which in turn are based on the FUER model. Following the coding and rating of these texts, the results of the groups were examined using methods of descriptive statistics. Moreover, the results at different stages and different groups were investigated via tests of comparison. For normally distributed data, *t*-tests were used for paired samples and ANOVA with repeated measures when comparing three points in time. For non-normal data, Wilcoxon signed-rank tests were performed for paired samples and Friedman tests when three samples were compared. The field notes taken during the implementation of the materials and the transcripts of the lessons were used to document the intervention and to corroborate the perceptions of the participants as voiced in the interviews. Thus, the field notes and lesson transcripts were only qualitatively analysed where relevant.

Following this, the thesis moved into presenting the empirical results. First, the process and the results of the pilot study were briefly outlined in chapter 6. Considering the amount of data of the main study, the reader was referred to Bauer-Marschallinger (2019), where the first pilot study is presented in more detail. In chapter 7, the results of the main study were presented, starting with the findings of the needs analyses, followed by an analysis of the design sessions. Here, the development of the materials was examined, and the materials produced were included and discussed as well. Then, the implementation phase was outlined before moving on to the evaluation of the interventions and the approach developed. In chapter 8, the empirical results were discussed in detail, structured according to research questions. Moreover, conceptual and methodological insights were discussed too. Key findings of the empirical part will now be presented in the following subchapter.

## 9.2 Key findings and implications

This subchapter presents key findings and implications in relation to the research questions that guided this study. It begins with presenting the design principles deduced from the empirical data and the literature reviewed (section 9.2.1), followed by the participants' responses to the intervention (9.2.2) and the effect of CDF-oriented teaching on learner language and content learning (9.2.3).

### 9.2.1 Design principles

The following presents a list of design principles produced in the course of this study. These specify the type and ideal characteristics of content-and-language-integrative materials and didactic procedures aimed at helping upper secondary CLIL history students improve their expression of cognitive processes when working on historical competences:

- 1) *Content-and-language-integrative materials should be student-centred, interactive, multi-modal, rich in variation, creative, and tailor-made resources that foster cognitive engagement.*
- 2) *Different scaffolding techniques, including a CDF-based approach, should be incorporated in content-and-language-integrative materials.*
- 3) *To genuinely integrate content and language, CDFs and form-meaning relations should be put into a subject-specific frame.*
- 4) *To genuinely integrate content and language in the subject history, central subject-specific concepts need to be considered, most notably contextualization and (retro-)perspectivity.*
- 5) *Upper secondary content-and-language-integrative history materials should address typical features of historical discourse, such as discipline-appropriate linking, hedging, and, to some degree, nominalisation.*
- 6) *Content-and-language-integrative materials should cater for mixed-ability groups to ensure an advantageous development for all learners.*

Starting with principle no. 1, the features listed – student-centredness, interaction, multi-modality, variation, creativity, custom-design, and cognitive engagement – match rather general quality criteria suggested in the literature. Learners and teachers, in line with CLIL researchers, call for materials that let learners get to work and keep them engaged via the inclusion of a variety of tasks and multi-modal input that require different interaction formats and ensure deep processing of the content.

Moving on to principle no. 2, participants of this study wished for more scaffolding ingrained in the materials they use in order to deal with the challenges of bilingual history education. The students appreciated general scaffolding techniques, like creating a progression of small, manageable steps, having glossaries, various ways of explanation, and peer support. This corresponds to the recommendations within the (CLIL) research community. More specifically, however, this research project has shown that the notion of CDFs works well as a basis for scaffolding historical input and competency-based tasks. First of all, they can be used to signpost

which cognitive operations are involved when deconstructing historical sources, reconstructing historical narratives, or just engaging with different forms of complex input through the creation of a manageable progression of small tasks. Secondly, explicit instruction on what prompts like “explain” or “evaluate” entail in the subject history counteracts the learners’ insecurities concerning performative verbs, which is in accordance with previous studies. Thirdly, as voiced by the participants and observable in their performances, providing learners with concrete tools and phrases to enact these communicative intentions helps learners verbalize their cognitive processes more precisely and appropriately within the discipline. Finally, in line with other studies, the findings of this study suggest that it is not only the learners’ understanding and realization of individual CDFs that need scaffolding but also the linking of several basic CDFs to sustain an overall communicative intention. This aspect, however, could not be fully captured in the materials developed in this study, considering that the different units only covered four to five lessons. In future projects, the combination and linking of individual CDFs into more substantial CDF episodes in subject-adequate ways, which brings us into the vicinity of genre-based teaching, should receive more attention.

Principle no. 3 deals with the transdisciplinary nature of content-and-language-integrative history teaching. Based on the results of this study, it appears that some learners and teachers only accept an overt focus on CDFs or language more generally if done explicitly through the lens of the discipline. In other words, when dealing with linguistic aspects, including CDFs, it makes sense to communicate how and why this is relevant within the subject. In one context of this study (A), form-focused teaching was rejected by the participants, while discussing form-meaning relations to express oneself in a subject-specific way was generally appreciated. In context B, the teacher did not consider accuracy and linguistic forms in the lessons, not even when the materials would have planned for that, reflecting the fact that he is not a language teacher (and does not want to be one). He did, however, pay attention to the construction of meaning. The learners of this group generally welcomed such a focus on language, but some would have rather welcomed more general, academic language input, insinuating that their EFL classes did not cover this aspect adequately. Overall, however, it seems that dealing with subject-specific language and CDF realizations in meaningful contexts is more in line with the curricular demands, the teacher’s self-concept, and in the majority of cases, the learners’ expectations.

Principle no. 4, then, zooms in on subject-specific considerations. Consistent with the history teaching literature, the learners of this study seemed unaware of certain subject-specific concepts or, at least, appeared to struggle with their application. For example, they often did not contextualize their claims and observations regarding historical sources and established facts. Here, explicit guidance on how to relate (COMPARE) historical knowledge (NARRATE/ REPORT) with the contents of historical sources (DESCRIBE) could help learners adhere to the conventions of the discipline. Moreover, the learners of this study often approached historical sources and content with a present-day bias, disregarding the historical context and zeitgeist. Additionally, they often

seemed unaware of the constructedness and subjectivity of historical sources, taking their contents at face value. Again, a step-by-step approach, thereby unfolding the complexity of 'doing history', could alleviate these issues.

Principle no. 5 is concerned with subject-specific linguistic features important for teaching CLIL history. Especially appropriate linking and hedging seemed to be important factors in the learners' demonstration of history skills, as their ratings correlated with content-related outcomes. Initially, these learners often failed to link their thoughts explicitly, thereby blurring their communicative intentions. In other cases, the cohesive devices used did not necessarily fit the function of the utterance or were rather repetitive, avoiding complex linkage and asyndetic structures typical for the subject history. Issues with hedging tied in with the learners' tendency to take the contents of historical sources at face value. Explicit attention to these features as provided by the intervention helped learners express their ideas more clearly and appropriately for the discipline, especially when deconstructing and evaluating historical visual sources. Nominalisation, on the other hand, was not a central issue initially and also did not correlate with content outcomes, contradicting previous assertions concerning the importance of nominalisations for historical discourse. It seems that more research is needed to establish the role of nominalisations in CLIL teaching. For now, the empirical results of this study suggest that nominalisations need not take priority in content-and-language-integrative CLIL (history) teaching. Yet, given its conceptual importance for constructing multi-factorial explanations, expressing collective agents as well as asyndetic structures, nominalisations may still need to receive some attention. However, this should not be covered in isolation but always in conjunction with the content.

Finally, principle no. 6 highlights the importance of differentiated instruction in CLIL. In the course of this study, it became apparent that catering to different levels of ability and work pace constitutes a central element for the success of the materials, both in terms of learner perceptions and actual outcomes. It turned out that taking into account different needs via providing a variety of task types, input material, and social formats did not suffice. Especially in the first cycle, learners complained about diverging paces and inappropriate levels of difficulty in both directions. We tried to counteract this development by including extra activities with more challenging tasks in later cycles, but in retrospect, more substantial differentiation strategies that do not run the risk of communicating a deficit model, where only gifted learners can 'shine' by doing interesting tasks, would be preferable. Nonetheless, the results of this study indicate that, by and large, weaker learners benefitted from the intervention more considerably than their high-achieving peers, reducing performance gaps, all in all. As such, the scaffolding strategies ingrained in the materials indeed are a first step to cater for the needs of weaker learners, counteracting CLIL's elitist connotation.

### 9.2.2 The participants' reactions to the intervention

In both contexts, the approach taken was markedly different from their traditional lessons. From the start, both teachers and learners appreciated the learner-centredness of the approach, reporting higher participation rates than usual. Learners and teachers also welcomed the type of tasks and their small steps. However, in both contexts, learners initially felt overwhelmed by the intensity of the lessons, highlighting the amount of writing, the dense programme, and the increased use of English. Moreover, focusing on subject-specific language felt unfamiliar and inappropriate at first. The teacher also took note of these issues in class. Reflecting on their own role, some of these points were relevant for the teachers too, such as the structural density and dealing with concerns of subject-specific and academic language in their content classes. Over the course of the project, however, the participants became accustomed to these changes, especially in context A, where two cycles took place.

Aside from this habituation effect, listening to the participants' feedback and adjusting the materials accordingly considerably increased their appreciation of the materials. For example, in the context of group A, making sure that explicit attention to language was always put into the perspective of the subject, prioritizing function over form, was positively received by the learners. Similarly, adding fast-track activities to counteract diverging paces and making sure that everyone feels challenged also significantly improved the learners' evaluation of the intervention.

As for perceived learning outcomes, most learners in both contexts felt that the intervention helped them learn the content of these units and, to a lesser extent, improve their language skills. While teacher A agreed with this estimation, T<sub>B2</sub> felt that the most tangible learning outcomes were of linguistic nature. In terms of educational value of individual tasks, the majority of tasks was well received, but those tasks where the language focus was not obviously linked to history education tended to be negatively assessed. The same was true for linguistic support measures. While there was no clear agreement on which type of language boxes were most helpful and which were redundant, those boxes that clearly communicated the connection between the linguistic features and the discipline tended to be positively evaluated. In turn, boxes containing more general and also simpler content were rejected by stronger students, who felt patronized by such measures. To avoid this, an easy solution would be to communicate to the learners the purpose of scaffolding and the possibility to ignore support measures one feels one does not need. For weaker students, on the other hand, the linguistic support measures turned out to be a crucial element to being able to cope with demanding tasks.

The overall structure of the materials was viewed positively by most participants. However, some stronger students felt limited by the structure of the materials, demanding more flexibility and autonomy. The teachers, however, did not share this view. They greatly appreciated the overall structure and the small steps, providing clear guidance, and argued that even their stronger students would benefit from clearer structures as provided in the intervention at hand. Nonetheless, an exploration of how further strategies of differentiation and individualisation can

be incorporated in future materials would be warranted. Moreover, as stressed by T<sub>B2</sub>, future materials should include more communicative and creative tasks too, preferably somewhere in the middle of the unit to keep learners involved and interested.

To summarize, throughout the project, an attitudinal shift could be observed in both contexts. In school A, learners were quite critical after the first round of intervention, but taking their feedback into account, paired with effects of habituation, resulted in a mostly positive evaluation of the materials after the second round of intervention. In context B, the learners expressed rather negative feelings about CLIL in general in their initial interview, but after the intervention, the interviewees seemed considerably more optimistic about CLIL and its potential, appreciating the materials developed in this study.

### 9.2.3 The effect of CDF-oriented teaching on the learners' historical competences and academic language skills

Starting with the learners' demonstration of their historical competences (methodological and orientation competence), comparison of pre- and post-intervention performances shows that the learners of group A improved considerably throughout the project as measured via the rubrics designed for the purpose of this study. After the first round of intervention, their results increased moderately, yet with statistical significance, followed by a strong, statistically significant leap in the final cycle. However, in group B, who started from a similar point of departure, average content results remained steady, presenting no statistical changes from pre- to post-intervention analysis.

Zooming in on the different dimensions under examination, both groups initially displayed similar strong and weak areas. At first, both groups struggled the most with justifying their claims comprehensibly and connecting their answers with the sources they were supposed to analyse (*justification/ comprehensibility*) as well as with including all expected elements in an appropriate amount of detail (*scope of content*). In the case of group A, these areas presented the sharpest growth overall, whereas in group B only minor, non-significant changes could be observed. In fact, results for *scope of content* even decreased in this group. It should also be noted that *justification/ comprehensibility* correlated most often with other descriptors, both in terms of content and language, indicating a central role of this aspect in historical discourse. Presenting answers in a systematic way (*systematicity*) was also a relatively weak area in the pre-intervention examinations. Here, group A increased significantly too but not as much as in other areas, with *systematicity* ranking second to last in the final results. Group B also only improved marginally and again to a statistically insignificant degree on the *systematicity* scale. Therefore, this would be an area needing more attention in future materials. Turning to the extent these learners presented accurate and/ or relevant points in their answers (*accuracy/ relevance*), results for this dimension started relatively low in both groups. In context A, this area increased moderately in cycle 1 and considerably in cycle 3, matching the overall trend of this cohort. Qualitatively, it could be

observed that the learners' initial present-day bias had faded somewhat, presenting answers that were considerably more sensitive towards historical contexts and which presented more accurate interpretations. In group B, slight improvements could be discerned yet again to a negligible extent, analogous to the overall development. Finally, the strongest areas in both contexts were *target competence* and *thinking level* at each point in time. In other words, at upper secondary level, learners seem to be relatively proficient in terms of demonstrating methodology competence and orientation competence at an appropriate level of historical thinking.

Turning to linguistic outcomes, average language results increased significantly from pre- to post-intervention examination as measured with linguistic rating rubrics developed for this study in both contexts. Again, the two groups started at similar levels and presented similar strong and weak areas. One of the weaker areas was *linking* both *in terms of form* and *function*, mostly because these learners rarely linked their ideas in the first place. After the intervention(s), this dimension presented the highest growth in both sets. In group A, results of these descriptors doubled throughout the project, and in group B statistically significant increases can be reported for *linking in terms of form*. In post-intervention performances, learners seemed capable of appropriately connecting their ideas, by and large, while also making use of a wider range of linking devices.

Scores for *hedging* started similarly low in both contexts. Unlike linking, however, results for *hedging* remained the lowest in the sets. Even so, some improvements can be reported. In context A, results rose to a statistically significant degree, whereas in group B there were only moderate and statistically insignificant increases. Qualitatively speaking, pre-intervention, most learners did not hedge their claims at all. Later, most learners tried to indicate caution linguistically more frequently, but they often failed to appropriately integrate these strategies into their own writing. Therefore, hedging would be an area needing more time and practice, ideally in conjunction with addressing the constructedness and subjectivity of historical sources.

Results for *nominalisation*, in contrast, started at a relatively high level in both groups. Again, in group A, scores increased significantly. Thus, this area remained one of the strongest in the set, with all but four students receiving the highest rating in the final post-intervention task. In group B, moderate improvements can be reported too, but yet again not with statistical significance. From a qualitative perspective, these learners indeed used a more nominal style after the intervention, but they also tended to make more mistakes.

Initial scores for *choice of CDF types* were the highest in each pre-intervention set. In the case of group A, significant increases can still be reported, while in group B only a negligible rise was observable. It appears that learners at upper secondary level are relatively proficient in choosing appropriate CDF types, but the intervention seemed to have had a positive effect nonetheless, at least in context A. Scores for *composition of CDF types*, in contrast, rose only moderately but to a statistically insignificant degree in context A. In context B, scores for this descriptor remained the same. This is especially problematic since a logical composition of CDF types seems to be a prerequisite for demonstrating history skills systematically and comprehensibly. When dealing

with complex tasks and input, logically assembling CDFs into a coherent episode is not an easy task, as has been demonstrated in other studies as well. For these reasons, more focused attention to the combination of CDFs seems warranted. In general, appropriate use of CDFs (choice and composition) appears to be a central element for meeting the requirements of competency-based tasks and ultimately for successfully demonstrating history skills. This claim is further backed up by statistical analyses displaying that these two CDF-related descriptors often correlated with *overall content* results. This, in turn, suggests that adequate CDF use is linked to adequate performance in the discipline.

### 9.3 Significance, limitations, and outlook

This thesis appears to be the first study that operationalized Dalton-Puffer's (2013) construct of cognitive discourse functions for upper secondary history CLIL education in order to blend content and language learning. While the importance of theorizing this integration has been well established (e.g., Nikula et al., 2016), relatively few attempts have been made at translating this line of research for classroom use (Donato, 2016; Meyer et al., 2015; Morton, 2020; Nashaat Sobhy, 2018). There are several reasons why such studies are warranted. First of all, teachers are in need of adequate materials which combine the perspective of the subject and the FL, and CLIL teachers also require pre- and in-service teacher training that prepares them for the challenges of content-and-language-integrative instruction, as voiced by the participants of this study but also as evident in the literature (Banegas, 2017; Gruber et al., 2020; Hahn, 2019; Meyer et al., 2015; Morton, 2013, 2020; Pérez Cañado, 2016a; van Kampen et al., 2018). Secondly, from the learners' perspective, research has shown that a 'language bath' is not enough to drive academic and subject-specific language skills, and thus a more purposeful integration is advised (Coyle et al., 2010; Gierlinger & Wagner, 2016; Meyer et al., 2015; Morton, 2010). Moreover, the demonstration of subject-specific skills, and therefore their performance in content subjects, seems contingent on the learners' skills to verbalize their cognitive operations when dealing with subject matter (e.g., Dalton-Puffer & Bauer-Marschallinger, 2019; Morton, 2020; Nashaat Sobhy, 2018; Nashaat-Sobhy & Llinares, 2020; Schleppegrell, 2004). What is more, some studies suggest that CLIL learners often feel frustration and exhaustion (Mearns et al., 2017; Ohlberger & Wegner, 2017; Rumlich, 2016), potentially because in- and output are not scaffolded enough (López-Medina, 2016; Morton, 2020; Otwinowska & Foryś, 2017; Somers & Llinares, 2018). Thirdly, such studies are needed from a conceptual point of view in order to empirically validate theoretical frameworks that integrate content and language learning. Dalton-Puffer (2013, 2016) and colleagues (Dalton-Puffer et al., 2018; Dalton-Puffer & Bauer-Marschallinger, 2019) have repeatedly called for classroom-based research that not only examines the relevance of the construct for teaching practice but also investigates its usefulness as a base for pedagogical planning and practice. The present study addresses these needs via a systematic, research-based development of pedagogical materials and design principles that operationalize the notion of



CDFs, aiming to combine the acquisition of subject-specific skills and the development of academic language.

For such an endeavour, it seems vital to transcend the boundaries of individual disciplines. In this case, perspectives of applied linguistics, history didactics, and general pedagogy were considered. Such a transdisciplinary perspective was facilitated by the background of the researcher, who is a trained teacher of English and history and who also taught CLIL history in the past. Looking at previous transdisciplinary research, it appears that little attention has been paid to the role of language within the field of history didactics (see chapter 4) compared to a considerable amount of linguistically-oriented studies into the connections of language and history education (see chapter 3). In general, the competency model underlying the Austrian history curriculum (FUER Geschichtsbewusstsein, Körber et al., 2007), though theoretically well described, has not been sufficiently examined from an empirical point of view (Kühberger, 2019), let alone in connection with aspects of (subject-specific) language learning. As such, the findings of this study might also be of interest to the field of history didactics.

Moreover, this study did not only go beyond the boundaries of the discipline but also of professional groups. To ensure that this study was both grounded in theory and practice-driven, a design-based research framework was adopted, which entails close collaboration between researcher and practitioners. Additionally, the interventions designed were trialled in natural testbeds, exploring different factors for the success of an intervention, which ultimately should allow the creation of ecologically valid and robust educational resources (McKenney & Reeves, 2012). In order to develop an approach that is accepted by learners too, this study also considered the voice of the target audience. Unfortunately, most CLIL design studies only consider the perspective of the students after the intervention. Yet, previous research suggests that learners are indeed capable of improving their own educational situation and that involving them in the design process increases the chance of a successful output (Coyle, 2013; Filice, 2021, see also Cook-Sather, 2006, 2020; Groundwater-Smith & Mockler, 2016; Mitra, 2018; Skinnari, 2020). The results of this study seem to confirm this, as listening to and incorporating the learners' feedback has not only positively impacted the learners' acceptance of the intervention on an affective level but also, presumably, in terms of learning outcomes, at least as measured in this study. This is illustrated by the fact that the learning curve of the learners in context A only appeared to really take off in the second round of intervention when their feedback had been incorporated in the materials, tying in with a much more positive evaluation by the learners. Overall, it seems fair to say that the intervention was successful. In each context, the learners' results improved significantly, with group B only improving in terms of language while group A, who received two treatments, presented significant gains in both domains. Moreover, on an affective level, participants agreed, by and large, that working with CDFs is an acceptable and useful way of combining content and language learning in the CLIL classroom, which also supports the learning process.

However, like most research, this project is subject to a number of limitations, many of which pertain to the exploratory character of this study and the limits of what one can do within one doctoral research project. Characteristically of DBR, this study aimed at examining “how we can make something work and why” (McKenney & Reeves, 2014, p. 143) rather than testing the efficiency of an intervention. In order to create a robust, applicable design while also developing context-sensitive design principles and theories about the inner workings of a pedagogical intervention, certain methodological decisions had to be made, which undoubtedly limit the generalizability of the results. To begin with, given the nature of the purpose and the context of this study, this project relied on qualitative methods and the quantification of interpretative analyses, which only allow for analytic and case-to-case generalization but not for statistical generalization to other populations. Nonetheless, the statistical analyses were still insightful, as they helped report and illustrate tendencies of the groups involved.

Likewise, the sample of this study had to be small to make such a project manageable. Moreover, a convenience sample was used for this study. Since teachers participating in this study had to devote considerable time and effort to this project too, only some few former colleagues volunteered, while teachers from other schools declined right away. However, I would like to highlight that the participating teachers were not only asked and selected because they seemed willing but also because they could bring valuable expertise to the project. As for the student sampling, although we could make some minor purposeful selections between parallel groups, the classes involved in this study were predetermined by the teachers’ allocation of classes. In other words, the participants of this study were selected on grounds of availability rather than a list of ideal criteria (e.g., non-selective CLIL context; representative groups in terms of academic achievement, motivation, or gender, etc.). Given these circumstances, there were contextual differences between the groups and teachers. However, as is typical for DBR, these variables were not muted since any contextual variances were potentially informative for the design.

Another issue present in this study, but which concerns DBR more generally, is that the intervention was evaluated by the same people who designed it, potentially amplifying the subjective element in this study. Considering that formative evaluations are more central than summative evaluations in DBR studies, this potential blurring of roles is tolerable. Nonetheless, the reports on the interventions’ effectiveness need to be treated with caution. This is especially true for the results of the written tests. As no validated research instruments for the assessment of the subject-specific competences or the CDF construct exist, two working-rubrics have been designed for the purpose of this study. Given the aims and the overall magnitude of the project, it was not possible to validate and reliably benchmark these tools. Such an endeavour would constitute a PhD thesis in its own right as, for example, is currently being conducted by del Pozo (in progress, see also del Pozo & Llinares, 2021). In general, more research into integrated ways of assessing learner performances is warranted (deBoer & Leontjev, 2020; Morton, 2020) so that we do not end up with two rubrics but one that truly embodies the content-and-language-

integrative nature of CLIL. The rubrics designed for this study might serve as an inspiration or starting point for such efforts, but they are definitely not the endpoint. While the intra-rater reliability scores are satisfactory, this, I assume, was only possible because I created a very detailed analysis manual over several rounds of analysis. Based on my insights during this process, I would recommend a higher level of granularity to allow for a more nuanced and clear-cut allocation of performances, and, like I mentioned above, a more genuinely integrated approach. Apart from creating useful research tools, such studies might consider the practical applicability of such tools, too, in order to support teachers in assessing CLIL learners; an area which many practitioners struggle with (see Morton, 2020; Otto & Estrada, 2019).

Additionally, I would like to point out that an inter-rater approach would have also yielded interesting insights into the reliability of the rubrics and the results reported in this study. Unfortunately, this was not feasible since no person was available that had both a background in history didactics and linguistics, being familiar with the FUER competences and CDFs at the same time. Alternatively, a second rater could have been recruited by offering extensive training, but this seemed unjustified for a mostly exploratory study. For the interview data, no intra- or inter-rating procedure was applied given the amount of data and complex (and different) coding schemes, with little potential added value for a qualitative content analysis, mainly of the structuring type (see Kuckartz, 2016; or Tedick & Young, 2018, who argued similarly).

Finally, it needs to be mentioned again that the data of this study was collected as part of the teachers' and learners' regular life at school. Naturally, this entailed a number of challenges that were not always foreseeable. While this might have affected the data collected to some extent, conducting this study in real-life and thus unpredictable classrooms enhanced the ecological validity of the design.

To account for all these complications and limitations, I purposefully employed a variety of research methods, and I considered several perspectives for the purpose of triangulation, as suggested by several DBR experts (e.g., Bakker & van Eerde; McKenney & Reeves, 2014). Moreover, following the recommendations of, for example, Euler (2014) or McKenney and Reeves (2012, 2014), I made sure to leave a thick audit trail as is hopefully visible throughout this thesis and the [appendix repository](#). I hope that, in this way, my findings and the process of reaching these insights are intersubjectively comprehensible and can therefore be of relevance for other researchers interested in the pedagogical operationalization of theoretical notions and/ or the integration of content and language learning. Nonetheless, in view of the local success of the intervention in this study, a large-scale, more quantitatively oriented study seems warranted to look into the general effectiveness of the approach. In addition, it would be interesting to further explore which dimensions of history skills and subject-specific language use are the most central factors for overall success, e.g., via regression analyses.

As regards the content of the intervention, the following limitations need to be addressed. First of all, the designs developed only focused on two of the four history competences. While these seem

to be the most essential ones in terms of overall objectives of the Austrian history curriculum and thus final history testing, other skills obviously do play a role in history teaching too. So, in future studies, it might make sense to also target other competences, and the same is true for CDFs. In this study, we centred on those CDF types that were most relevant for the competences in focus, most notably EVALUATE. Future research should pay more attention to the other types too (both in terms of operationalization and testing), e.g., EXPLAIN in connection with reconstruction/ narrative competence or COMPARE together with terminology competence. Moreover, as has been mentioned in the previous subchapter, future studies should pay more attention to how several CDFs can be reasonably combined in a subject-appropriate way. The L2 users in this and other studies (Breeze & Dafouz, 2017; Doiz & Lasagabaster, 2021) struggled with this aspect while, at the same time, a logical, discipline-appropriate composition of CDF types appears vital for demonstrating subject-specific skills competently. Unfortunately, this insight only crystallized in the process of the study and would therefore need more attention from the start in future design studies.

Additionally, considering the diversity of learners more substantially than was possible in this study is advised. As the results of this project have gradually revealed, catering to different needs and levels of ability and pace turned out to be crucial for the participants' response to the intervention. Given that this project was conducted in streamed CLIL programmes, suggesting a rather homogenous sample, it is all the more important to consider this aspect in contexts where CLIL learners are not selected as is the case in many school types all over Europe now. Interesting questions would be what learners in diverse and unstreamed CLIL settings need beyond catering to different paces and ability levels, and what type of strategies for scaffolding and differentiated instruction would achieve most optimal results for different groups of learners in such settings.

Reflecting on the various elements of this research project, I would like to address some factors or variables that played an important role in the design and its implementation but could not be investigated more thoroughly within the constraints of one, though rather large, research project. For example, considerations concerning the learners' affective factors definitely co-determined the design process, albeit rather implicitly, as this was not a central theoretical factor in the study at hand. From the teachers' perspective, this was one of the central aspects and, in case of teacher A, one of the main reasons for participating in this project. Given the importance of motivation both subjectively and in relation to learning outcomes (Lasagabaster, 2011; Wesely, 2012), future design research should consider motivation in a more nuanced way, including reaching a better understanding of what motivation means in CLIL conceptually. This might entail examining the role of an integrated CLIL motivation rather than motivation towards the FL, as Somers and Llinares (2018) have suggested.

Another interesting aspect is the role of learner discourse and translanguaging. Given the amount of data this thesis elicited, the transcripts of the lessons could not be analysed comprehensively. However, these would provide valuable insights into how learners enact these CDFs when preparing for outputs, i.e., all the group and pair work phases or teacher-student-talk. Interesting

questions would be to what extent these processes are multilingual and co-constructed. While some initial observations were made in this thesis, a more thorough analysis would yield more substantial findings concerning the linguistic practices of CLIL learners and teachers and how different linguistic resources are being put to use in CLIL classrooms. This could further inform CLIL materials and content-and-language-integrative approaches.

Looking at the context of this study, the materials and design principles created are tailored to upper secondary, rather advanced CLIL learners. However, it would be interesting how such an approach would work on other educational levels and in other subjects. Some initial work has already been started or even completed, especially in history on tertiary level (Doiz & Lasagabaster, 2021; Lasagabaster et al., 2021) or science on secondary level (Breeze & Gerns, 2019; Connolly, 2019; Hasenberger, 2018), but further work is needed to create a more comprehensive approach to a pedagogical integration of content and language learning. This also means targeting a younger audience to ensure ideal starting conditions, which then open up possibilities for continuous learning trajectories of subject-literacy skills. From a practical point of view, teachers need more materials to draw from and opportunities to further develop in this regard via in-service teacher training. While these materials and design principles are set up in a way that facilitates adaptations to other topics, this still requires temporal resources on the teachers' part, which might not always be available.

To conclude, this study contributes to helping CLIL teachers and their learners to better cope with the challenges of CLIL history education. However, as this last subchapter has shown, there are still many blind spots and under-researched areas in this regard, especially when it comes to practice-oriented yet theoretically grounded research. This type of "translational" research is needed if we want to see research-based innovations put to life in class. I hope that future research will further strengthen the links between practice and research so that these two perspectives can continue to drive one another. This way, CLIL education might be able to really reach its full potential, helping learners gain a voice in their subject.

## References

- Abello-Contesse, C. (2013). Bilingual and multilingual education: An overview of the field. In C. Abello-Contesse, P. M. Chandler, M. D. López-Jiménez, & R. Chacón-Beltrán (Eds.), *Bilingual and multilingual education in the 21st century* (pp. 3–23). Multilingual Matters.
- Achugar, M., & Carpenter, B. D. (2012). Developing disciplinary literacy in a multilingual history classroom. *Linguistics and Education*, 23(3), 262–276.  
<https://doi.org/10.1016/j.linged.2012.05.003>
- Achugar, M., & Schleppegrell, M. J. (2005). Beyond connectors: The construction of cause in history textbooks. *Linguistics and Education*, 16(3), 298–318.  
<https://doi.org/10.1016/j.linged.2006.02.003>
- Achugar, M., Schleppegrell, M. J., & Oteiza, T. (2007). Engaging teachers in language analysis: A functional linguistics approach to reflective literacy. *English Teaching: Practice and Critique*, 6(2), 8–22.
- ADiBE. (2021). *Attention to diversity in CLIL (Erasmus+ project)*. University of Jaen.  
<https://adibeproject.com/>
- Admiraal, W., Westhoff, G., & Bot, K. de (2006). Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English. *Educational Research and Evaluation*, 12(1), 75–93. <https://doi.org/10.1080/13803610500392160>
- Agustín-Llach, M. P. (2015). The effects of the CLIL approach in young foreign language learners' lexical profiles. *International Journal of Bilingual Education and Bilingualism*, 20(5), 557–573. <https://doi.org/10.1080/13670050.2015.1103208>
- Agustín-Llach, M. P. (2016). Age and type of instruction (CLIL vs. traditional EFL) in lexical development. *International Journal of English Studies*, 16(1), 75–96.
- Agustín-Llach, M. P., & Canga Alonso, A. (2016). Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education. *International Journal of Applied Linguistics*, 26(2), 211–227. <https://doi.org/10.1111/ijal.12090>
- Agustín-Llach, M. P., & Jiménez-Catalán, R. M. (2018). Teasing out the role of age and exposure in EFL learners' lexical profiles: A comparison of children and adults. *International Review of Applied Linguistics in Language Teaching*, 56(1), 25–43.
- Ahern, A., Whittaker, R., & Sánchez, I. B. (2018). Reading and writing to learn: A principled approach to practice in CLIL/bilingual classes. *E-TEALS*, 9(1), 23–40.  
<https://doi.org/10.2478/eteals-2018-0011>
- Albert, R., & Marx, N. (2016). *Empirisches Arbeiten in Linguistik und Sprachlehrforschung [empirical research in linguistics and language teaching research]: Anleitung zu quantitativen Studien von der Planungsphase bis zum Forschungsbericht [a guide to quantitative studies from planning to reporting results]* (3rd rev. ed.). Narr Studienbücher. Narr Francke Attempto.
- Alejo, R., & Piquer-Píriz, A. (2016). Urban vs. rural CLIL: An analysis of input-related variables, motivation and language attainment. *Language, Culture and Curriculum*, 29(3), 245–262.  
<https://doi.org/10.1080/07908318.2016.1154068>
- Alonso, E., Grisaleña, J., & Campo, A. (2008). Plurilingual education in secondary schools: Analysis of results. *International CLIL Research Journal*, 1(1), 36–49.

- Ament, J. R., & Pérez-Vidal, C. (2015). Linguistic outcomes of English Medium Instruction programmes in higher education: A study on economics undergraduates at a Catalan university. *Higher Learning Research Communications*, 5(1), 47.  
<https://doi.org/10.18870/hlrc.v5i1.239>
- Ammerer, H., & Windischbauer, E. (2011). *Kompetenzorientierter Unterricht in Geschichte und Politischer Bildung [competency-based teaching in history and political education]: Diagnoseaufgaben mit Bildern [diagnostic tasks with visuals]*. Edition polis.
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Longman.
- Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, 41(1), 16–25.  
<https://doi.org/10.3102/0013189X11428813>
- Arnould, G. L. W. (1876). The Burbach smelting works near Saarbrücken, Bildarchiv Preußischer Kulturbesitz. [https://germanhistorydocs.ghi-dc.org/sub\\_image\\_s.cfm?image\\_id=1333&language=english](https://germanhistorydocs.ghi-dc.org/sub_image_s.cfm?image_id=1333&language=english)
- Arribas, M. (2016). Analysing a whole CLIL school: Students' attitudes, motivation, and receptive vocabulary outcomes. *Latin American Journal of Content and Language Integrated Learning*, 9(2), 267–292. <https://doi.org/10.5294/lacil.2016.9.2.2>
- Artieda, G., Roquet, H., & Nicolás-Conesa, F. (2017). The impact of age and exposure on EFL achievement in two learning contexts: Formal instruction and formal instruction + content and language integrated learning (CLIL). *International Journal of Bilingual Education and Bilingualism*, 9(4), 1–19.  
<https://doi.org/10.1080/13670050.2017.1373059>
- Austin, J. L. (Ed.). (1962). *How to do things with words*. Harvard Univ. Press.
- Austrian Federal Ministry for Education. (1995). *Lehrplan für allgemeinbildende höhere Schulen [curriculum for academic secondary schools]: BGBl. no. 88/1985*.  
<https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568>
- Austrian Federal Ministry for Education. (2004). *Lehrpläne für allgemeinbildende höhere Schulen [curriculum for academic secondary schools]: BGBl. no. 88/1985*.  
<https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568>
- Austrian Federal Ministry for Education. (2005). *Lehrplan der Volksschule [curriculum of primary schools]: BGBl. II no. 368/2005*.  
<https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20009288>
- Austrian Federal Ministry for Education. (2011). *Die kompetenzorientierte mündliche Reifeprüfung [the competency-based oral final exam]: Geschichte [history]*.  
[https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp\\_ahs/mrp\\_flf.html](https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp_ahs/mrp_flf.html)
- Austrian Federal Ministry for Education. (2013). *Schulartenübergreifender Bildungsstandard in der Berufsbildung [educational standards in different vocational schools]: Geografie, Geschichte und politischer Bildung einschl. Volkswirtschaftlicher Grundlagen [geography,*

- history, political education incl. economic basics*.  
<https://www.bildungsstandards.berufsbildendeschulen.at/bildungsstandards/schulartenuebergreifende-bildungsstandards>
- Austrian Federal Ministry for Education. (2014). *Curriculum for the secondary college of business administration: BGBl. no. 895/1994*. <https://www.hak.cc/node/4336>
- Austrian Federal Ministry for Education. (2015a). *Lehrplan der Höheren Technischen Lehranstalt [curriculum of the higher federal technical college]: BGBl. II no. 262/2015*.  
<https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20009288>
- Austrian Federal Ministry for Education. (2015b). *Lehrplan Höhere Lehranstalt für Mode [curriculum for the secondary school for fashion]: BGBl. II no. 340/2015*.  
<https://www.abc.berufsbildendeschulen.at/downloads/?kategorie=14>
- Austrian Federal Ministry for Education. (2015c). *Lehrplan Höhrere Lehranstalt für Tourismus [curriculum for the secondary school for tourism]: BGBl. II no. 340/2015*.  
<https://www.abc.berufsbildendeschulen.at/downloads/?kategorie=14>
- Austrian Federal Ministry for Education. (2017a). *The Austrian education system*.  
[https://bildung.bmbwf.gv.at/enfr/school/bw\\_en/bildungswege2016\\_eng.pdf?61ec3r](https://bildung.bmbwf.gv.at/enfr/school/bw_en/bildungswege2016_eng.pdf?61ec3r)
- Austrian Federal Ministry for Education. (2017b). *CLIL: Handreichung zur Umsetzung von CLIL an höheren land- und forstwirtschaftlichen Schulen [guidelines for the implementation of CLIL at secondary colleges for agriculture and forestry]*.  
[https://www.hum.at/images/unterrichtsentwicklung/CLIL/Leitfaden\\_HLFS\\_06112017\\_Version2.pdf](https://www.hum.at/images/unterrichtsentwicklung/CLIL/Leitfaden_HLFS_06112017_Version2.pdf)
- Austrian Federal Ministry for Education. (2018a). *Lehrplan der Bildungsanstalt für Elementarpädagogik [curriculum of the secondary college for early childhood pedagogy]: BGBl. I no. 103/2018*.  
<https://www.abc.berufsbildendeschulen.at/downloads/?kategorie=14>
- Austrian Federal Ministry for Education. (2018b). *Lehrplan der Mittelschulen [curriculum of the compulsory secondary school]: BGBl. II no. 185/2012*.  
<https://www.bmbwf.gv.at/Themen/schule/schulpraxis/lp.html>
- Austrian Federal Ministry for Education. (2019). *Aufnahme in eine allgemein bildende höhere Schule (AHS) [admission to academic secondary schools (AHS)]*.  
[https://www.bmbwf.gv.at/Themen/schule/beratung/schulinfo/aufnahme\\_ahs.html](https://www.bmbwf.gv.at/Themen/schule/beratung/schulinfo/aufnahme_ahs.html)
- Badertscher, H., & Bieri, T. (2009). *Wissenserwerb im Content and Language Integrated Learning [knowledge acquisition in content and language integrated learning]: Empirische Befunde und Interpretationen [empirical evidence and interpretations]* (1st ed.). Haupt.
- Baek, J. Y., & Bannan-Ritland, B. (2008). Investigating the act of design in design research. In A. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 299–319). Routledge.
- Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 Education: A Design Document: CSE Report 611*. Los Angeles. National Center for Research on Evaluation, Standards, and Student Testing University of California, Los Angeles.



- Bailey, A. L., Huang, B., Shin, H. W., Farnsworth, T., & Butler, F. A. (2007). *Developing academic English language proficiency prototypes for 5th grade reading: Psychometric and linguistic profiles of tasks: An extended executive summary*. Los Angeles. National Center for Research on Evaluation, Standards, and Student Testing University of California, Los Angeles.
- Bakker, A., & van Eerde, D. An introduction to design-based research with an example from statistics education. In A. Bikner-Ahsbals, C. Knipping, & N. C. Presmeg (Eds.), *Advances in mathematics education. Approaches to qualitative research in mathematics education: Examples of methodology and methods* (pp. 429–466). Springer.
- Ball, P. (2018). Innovations and challenges in CLIL materials design. *Theory into Practice*, 57(3), 222–231. <https://doi.org/10.1080/00405841.2018.1484036>
- Ball, P., Kelly, K., & Clegg, J. (2015). *Putting CLIL into practice. Oxford handbooks for language teachers*. Oxford University Press.
- Banegas, D. L. (2012). CLIL teacher development: Challenges and experiences. *Latin American Journal of Content and Language Integrated Learning*, 5(1), 46–56. <https://doi.org/10.5294/lacil.2012.5.1.4>
- Banegas, D. L. (2013). *Teachers developing language-driven CLIL through collaborative action research in Argentina* [PhD thesis]. The University of Warwick, Warwick.
- Banegas, D. L. (2014). An investigation into CLIL-related sections of EFL coursebooks: Issues of CLIL inclusion in the publishing market. *International Journal of Bilingual Education and Bilingualism*, 17(3), 345–359. <https://doi.org/10.1080/13670050.2013.793651>
- Banegas, D. L. (2017). Teacher-developed materials for CLIL: Frameworks, sources, and activities. *Asian EFL Journal*, 19(3), 31–48.
- Banegas, D. L., & Pinner, R. (2021). Motivations and synergy on a sociolinguistics module in language teacher education in Argentina. In K. R. Talbot, M.-T. Gruber, & R. Nishida (Eds.), *The psychological experience of integrating content and language* (pp. 266–282). Multilingual Matters.
- Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, 13(1), 1–14. [https://doi.org/10.1207/s15327809jls1301\\_1](https://doi.org/10.1207/s15327809jls1301_1)
- Barricelli, M., Gautschi, P., & Körber, A. (2012). Historische Kompetenzen und Kompetenzmodelle [Historical competences and competency-models]. In M. Barricelli & M. Lücke (Eds.), *Handbuch Praxis des Geschichtsunterrichts [handbook of the practice of history teaching]* (pp. 207–235). Wochenschau Verlag.
- Barrios, E., & Milla Lara, M. D. (2020). CLIL methodology, materials and resources, and assessment in a monolingual context: An analysis of stakeholders' perceptions in Andalusia. *The Language Learning Journal*, 48(1), 60–80. <https://doi.org/10.1080/09571736.2018.1544269>
- Bartlett, T. (2017). Context in systemic functional linguistics: Towards scalar supervenience? In T. Bartlett & G. O'Grady (Eds.), *Routledge handbooks. The Routledge handbook of systemic functional linguistics* (pp. 375–390). Routledge.
- Barwell, R. (2016). A Bakhtinian perspective on language and content integration: Encountering the alien word in second language mathematics. In T. Nikula, E. Dafouz, P. Moore, & U.

- Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (pp. 101–122). Multilingual Matters.
- Bauer-Marschallinger, S. (2016). *Acquisition of historical competences in the CLIL history classroom* [diploma thesis]. University of Vienna, Vienna.  
<https://theses.univie.ac.at/detail/37216>
- Bauer-Marschallinger, S. (2019). With united forces: How design-based research can link theory and practice in the transdisciplinary sphere of CLIL. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, 2(2), 7–23.  
<https://doi.org/10.5565/rev/clil.19>
- Bauer-Marschallinger, S., Dalton-Puffer, C., Heaney, H., Katzinger, L., & Smit, U. (2021). CLIL for all? An exploratory study of reported pedagogical practices in Austrian secondary schools. *International Journal of Bilingual Education and Bilingualism* (ahead of print).  
<https://doi.org/10.1080/13670050.2021.1996533>
- Bayram, D., Öztürk, R. Ö., & Atay, D. (2019). Reading comprehension and vocabulary size of CLIL and non-CLIL students: A comparative study. *Language Teaching and Educational Research*, 101–113. <https://doi.org/10.35207/late.639337>
- Beacco, J.-C. (2010). *Items for a description of linguistic competence in the language of schooling necessary for learning/teaching history (end of obligatory education): An approach with reference points*. Language Policy Division. Directorate of Education and Languages.
- Beacco, J.-C., Coste, D., van de Ven, P.-H., & Vollmer, H. J. (2010). *Language and school subjects: Linguistic dimensions of knowledge building in school curricula*. Language Policy Division. Directorate of Education and Languages.
- Berelson, B. (1952). *Content analysis in communication research*. Free press.
- Bernstein, B. (1999). Vertical and horizontal discourse: An essay. *British Journal of Sociology of Education*, 20(2), 157–173. <https://doi.org/10.1080/01425699995380>
- Bernstein, B. (2003). *The structuring of pedagogic discourse. Class, codes, and control: Vol. 3*. Routledge.
- Biggs, J. B., & Tang, C. (2011). *Teaching for quality learning at university: What the student does* (4. ed.). *SRHE and Open University Press imprint*. McGraw-Hill, Society for Research into Higher Education & Open University Press.
- Bloom, B. S., Engelhart, M., Furst, E., Jill, W., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans, Green.
- Bohnsack, R. (2010). Group discussion and focus groups. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 214–221). SAGE Publications.
- Bonnet, A., & Breidbach, S. (2017). CLIL teachers' professionalization: Between explicit knowledge and professional identity. In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 269–287). John Benjamins.
- Bortz, J., & Döring, N. (2016). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaften [research methods and evaluation for the humanities and social sciences]* (5th rev. ed.). *Springer-Lehrbuch*. Springer.

- Breeze, R., & Dafouz, E. (2017). Constructing complex cognitive discourse functions in higher education: An exploratory study of exam answers in Spanish- and English-medium instruction settings. *System*, 70, 81–91. <https://doi.org/10.1016/j.system.2017.09.024>
- Breeze, R., & Gerns, P. (2019). Building literacies in secondary school history: The specific contribution of academic writing support. *EuroAmerican Journal of Applied Linguistics and Languages*, 6(1), 21–36. <https://doi.org/10.21283/2376905X.10.149>
- Broca, Á. (2016). CLIL and non-CLIL: Differences from the outset. *ELT Journal*, 70(3), 320–331. <https://doi.org/10.1093/elt/ccw011>
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex Interventions in classroom settings. *Journal of the Learning Sciences*, 2(2), 141–178. [https://doi.org/10.1207/s15327809jls0202\\_2](https://doi.org/10.1207/s15327809jls0202_2)
- Bruner, J. S. (1960). *The process of education*. Harvard University Press. <https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=3300117>
- Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, 39(4), 523–532. <https://doi.org/10.1016/j.system.2011.08.002>
- Bruton, A. (2013). CLIL: Some of the reasons why ... and why not. *System*, 41(3), 587–597. <https://doi.org/10.1016/j.system.2013.07.001>
- Bruton, A. (2017). Questions about CLIL which are unfortunately still not outdated: A reply to Pérez-Cañado. *Applied Linguistics Review*, 10(4), 591–602. <https://doi.org/10.1515/applirev-2017-0059>
- Bulon, A. (2020). Comparing the ‘phrasicon’ of teenagers in immersive and non-immersive settings: Does input quantity impact range and accuracy? *Journal of Immersion and Content-Based Language Education*, 8(1), 107–136.
- Burwitz-Melzer, E., & Steininger, I. (2016). Inhaltsanalyse [content analysis]. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik [research methods in language teaching research]: Ein Handbuch [a handbook]* (1st ed., pp. 256–268). Narr Francke Attempto.
- Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L. (2004). *An approach to operationalizing academic language for language test development purposes: Evidence from fifth-grade science and math*. Los Angeles.
- Cabezuelo Gutierrez, P., & Fernández, R. (2014). A case study on teacher training needs in the Madrid bilingual project. *Latin American Journal of Content and Language Integrated Learning*, 7(2), 49–70. <https://doi.org/10.5294/laclil.2014.7.2.3>
- Calderón-Jurado, B., & Garcia, C. M. (2018). Students’ attitude and motivation in bilingual education. *International Journal of Educational Psychology*, 7(3), 317. <https://doi.org/10.17583/ijep.2018.3558>
- Cammarata, L., & Cavanagh, M. (2018). Teacher education and professional development for immersion and content-based instruction. *Journal of Immersion and Content-Based Language Education*, 6(2), 189–217. <https://doi.org/10.1075/jicb.18009.cam>
- Cammarata, L., Tedick, D. J., & Osborn, T. A. (2016). Content-based instruction and curricular reforms: Issues and goals. In L. Cammarata (Ed.), *Content-based foreign language teaching: Curriculum and pedagogy for developing advanced thinking and literacy skills* (pp. 1–22). Routledge.

- Canga Alonso, A. (2015a). Receptive vocabulary of CLIL and non-CLIL primary and secondary school learners. *Complutense Journal of English Studies*, 23, 59–77.
- Canga Alonso, A. (2015b). The receptive vocabulary size of Spanish 5th grade primary school students in CLIL and non-CLIL instruction. *ES*, 36, 63–85.
- Carpenter, B. D., Earhart, M., & Achugar, M. (2014). Working with documents to develop disciplinary literacy in the multilingual classroom. *The History Teacher*, 48(1), 91–103.
- Carretero, M., & van Alphen, F. (2014). Do master narratives change among high school students? A characterization of how national history is represented. *Cognition and Instruction*, 32(3), 290–312. <https://doi.org/10.1080/07370008.2014.919298>
- Caspari, D. (2016). Grundfragen fremdsprachendidaktischer Forschung [foundations of language didactics research]. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik [research methods in language teaching research]: Ein Handbuch [a handbook]* (1st ed., pp. 141–153). Narr Francke Attempto.
- Castellano-Risco, I. (2018). Receptive vocabulary and learning strategies in secondary school CLIL and non-CLIL learners. *Onomázein Revista De Lingüística Filología Y Traducción*(40), 28–48. <https://doi.org/10.7764/onomazein.40.02>
- Castellano-Risco, I., Alejo-González, R., & Piquer-Píriz, A. M. (2020). The development of receptive vocabulary in CLIL vs EFL: Is the learning context the main variable? *System*, 91, 102263. <https://doi.org/10.1016/j.system.2020.102263>
- Castro-García, D. (2017). Receptive vocabulary measures for EFL Costa Rican high school students. *International Journal of English Studies*, 17(2), 81–99.
- Cenoz, J., Genesee, F., & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *Applied Linguistics*, 35(3), 243–262. <https://doi.org/10.1093/applin/amt011>
- Christie, F., & Derewianka, B. (2008). *School discourse: Learning to write across the years of schooling*. Continuum discourse series. Continuum.
- Cleve, J., & Lämmel, U. (2016). *Data mining* (2nd ed.). De Gruyter Oldenbourg.
- Codina-Espurz, V., & Salazar-Campillo, P. (2019). Openings and closing in emails by CLIL students: A pedagogical proposal. *English Language Teaching*, 12(2), 57–67.
- Coffin, C. (2006). *Historical discourse: The language of time, cause, and evaluation*. Continuum discourse series. Continuum.
- Coffin, C. (2017). Introduction to part II: Systemic Functional Linguistics: A theory for integrating content-language learning (CLIL). In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 91–104). John Benjamins.
- Coffin, C., Acevedo, C., Löfstedt, A.-C., & Rose, D. (2013). *Teacher learning for European literacy education (TeL4ELE): Final report public part*. <https://doi.org/10.13140/RG.2.2.26758.29764>
- Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Routledge.
- Connolly, T. (2019). *Die Förderung vertiefter Lernprozesse durch Sachfachliteralität [supporting deeper learning processes through subject-literacy]: Eine vergleichende Studie zum expliziten Scaffolding kognitiver Diskursfunktionen im bilingualen Chemieunterricht am*

- Beispiel des Erklärens [a comparative study on explicit scaffolding of cognitive discourse functions in bilingual chemistry education using the example of explaining]* [PhD thesis]. Johannes Gutenberg-Universität Mainz, Mainz.
- Cook-Sather, A. (2006). Sound, presence, and power: "Student voice" in educational research and reform. *Curriculum Inquiry*, 36(4), 359–390.
- Cook-Sather, A. (2020). Student voice across contexts: Fostering student agency in today's schools. *Theory into Practice*, 59(2), 182–191.  
<https://doi.org/10.1080/00405841.2019.1705091>
- Coyle, D. (2013). Listening to learners: An investigation into 'successful learning' across CLIL contexts. *International Journal of Bilingual Education and Bilingualism*, 16(3), 244–266.  
<https://doi.org/10.1080/13670050.2013.777384>
- Coyle, D., Bower, K., Foley, Y., & Hancock, J. (2021). Teachers as designers of learning in diverse bilingual classrooms: A UK case study. *International Journal of Bilingual Education and Bilingualism*. <https://doi.org/10.1080/13670050.2021.1989373>
- Coyle, D., Halbach, A., Meyer, O., & Schuck, K. (2018). Knowledge ecology for conceptual growth: Teachers as active agents in developing a pluriliteracies approach to teaching for learning (PTL). *International Journal of Bilingual Education and Bilingualism*, 21(3), 349–365. <https://doi.org/10.1080/13670050.2017.1387516>
- Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge University Press.
- Crossman, K. (2018). Immersed in academic English: Vocabulary and academic outcomes of a CLIL university preparation course. *International Journal of Bilingual Education and Bilingualism*, 21(5), 564–577. <https://doi.org/10.1080/13670050.2018.1494698>
- Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, 14(2), 175.  
<https://doi.org/10.2307/3586312>
- Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy* (Vol. 6). Multilingual Matters.
- Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. V. Street & N. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 71–83). Springer. [https://doi.org/10.1007/978-0-387-30424-3\\_36](https://doi.org/10.1007/978-0-387-30424-3_36)
- Cummins, J. (2017). BICS and CALP: Empirical and theoretical status of the distinction. In B. V. Street & S. May (Eds.), *Literacies and language education* (3rd ed., Vol. 62, pp. 59–71). Springer. [https://doi.org/10.1007/978-3-319-02252-9\\_6](https://doi.org/10.1007/978-3-319-02252-9_6)
- Dafouz, E., Camacho, M., & Urquia, E. (2014). 'Surely they can't do as well': A comparison of business students' academic performance in English-medium and Spanish-as-first-language-medium programmes. *Language and Education*, 28(3), 223–236.  
<https://doi.org/10.1080/09500782.2013.808661>
- Dallinger, S., Jonkmann, K., & Hollm, J. (2018). Selectivity of content and language integrated learning programmes in German secondary schools. *International Journal of Bilingual Education and Bilingualism*, 21(1), 93–104.  
<https://doi.org/10.1080/13670050.2015.1130015>

- Dallinger, S., Jonkmann, K., Hollm, J., & Fiege, C. (2016). The effect of content and language integrated learning on students' English and history competences: Killing two birds with one stone? *Learning and Instruction*, 41, 23–31.  
<https://doi.org/10.1016/j.learninstruc.2015.09.003>
- Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms. Language learning and language teaching: Vol. 20*. John Benjamins.
- Dalton-Puffer, C. (2008). Outcomes and processes in content and language integrated learning (CLIL): Current research from Europe. In W. Delanoy (Ed.), *Anglistische Forschungen. Future perspectives for English language teaching* (Vol. 388, pp. 139–157). Winter.
- Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, 31, 182–204.  
<https://doi.org/10.1017/S0267190511000092>
- Dalton-Puffer, C. (2013). A construct of cognitive discourse functions for conceptualising content-language integration in CLIL and multilingual education. *European Journal of Applied Linguistics*, 1(2), 99. <https://doi.org/10.1515/eujal-2013-0011>
- Dalton-Puffer, C. (2016). Cognitive discourse functions: Specifying and integrative interdisciplinary construct. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (pp. 29–54). Multilingual Matters.
- Dalton-Puffer, C. (2017). Same but different: Content and language integrated learning and content-based instruction. In M. A. Snow & D. Brinton (Eds.), *Michigan teacher training. The content-based classroom: New perspectives on integrating language and content* (pp. 151–164). University of Michigan Press.
- Dalton-Puffer, C. (2018). Postscriptum: Research pathways in CLIL/ immersion instructional practices and teacher development. *International Journal of Bilingual Education and Bilingualism*, 21(3), 384–387. <https://doi.org/10.1080/13670050.2017.1384448>
- Dalton-Puffer, C., & Bauer-Marschallinger, S. (2019). Cognitive discourse functions meet historical competences: Towards an integrated pedagogy in CLIL history education. *Journal of Immersion and Content-Based Language Education*, 7(1), 30–60.  
<https://doi.org/10.1075/jicb.17017.dal>
- Dalton-Puffer, C., Bauer-Marschallinger, S., Brückl-Mackey, K., Hofmann, V., Hopf, J., Kröss, L., & Lechner, L. (2018). Cognitive discourse functions in Austrian CLIL lessons: Towards an empirical validation of the CDF construct. *European Journal of Applied Linguistics*, 6(1), 5–29. <https://doi.org/10.1515/eujal-2017-0028>
- Dalton-Puffer, C., Faistauer, R., & Vetter, E. (2011). Research on language teaching and learning in Austria (2004–2009). *Language Teaching*, 44(2), 181–211.  
<https://doi.org/10.1017/S0261444810000418>
- Dalton-Puffer, C., Hüttner, J., & Smit, U. (2021). From voluntary to obligatory CLIL in upper secondary technical colleges: Teacher and student voices from a diverse landscape. In K. R. Talbot, M.-T. Gruber, & R. Nishida (Eds.), *The psychological experience of integrating content and language* (pp. 93–112). Multilingual Matters.

- Dalton-Puffer, C., & Llinares, A. G. (2015). The role of different tasks in CLIL students' use of evaluative language. *System. Special Issue: The Interface Between Task-Based Language Teaching and Content-Based Instruction*, 54, 69–79.
- Dalton-Puffer, C., Llinares, A. G., Lorenzo, F., & Nikula, T. (2014). "You can stand under my umbrella": Immersion, CLIL and bilingual education: A response to Cenoz, Genesee & Gorter (2013). *Applied Linguistics*, 35(2), 213–218.  
<https://doi.org/10.1093/applin/amu010>
- Dalton-Puffer, C., Nikula, T., & Smit, U. (Eds.). (2010). *AILA applied linguistics series. Language use and language learning in CLIL classrooms*. John Benjamins.
- de Graaff, R. (2016). Integrating content and language in education: Best of both worlds? Foreword. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (xiii-xvu). Multilingual Matters.
- de Graaff, R., Koopman, G. J., Anikina, Y., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in content and language integrated learning (CLIL). *International Journal of Bilingual Education and Bilingualism*, 10(5), 603–624.  
<https://doi.org/10.2167/beb462.0>
- de Graaff, R., Koopman, G. J., & Westhoff, G. (2007). Identifying effective L2 pedagogy in content and language integrated learning (CLIL). *VIEWS: Vienna English Working Papers*, 16(3), 12–19.  
[http://anglistik.univie.ac.at/fileadmin/user\\_upload/dep\\_anglist/weitere\\_Uploads/Views/Views\\_0703.pdf](http://anglistik.univie.ac.at/fileadmin/user_upload/dep_anglist/weitere_Uploads/Views/Views_0703.pdf)
- de Oliveira, L. (2010). Nouns in history: Packaging information, expanding explanations, and structuring reasoning. *The History Teacher*, 43(2), 191–203.
- deBoer, M. (2020). Teacher-based assessment of learner-led interactions in CLIL: The power of cognitive discourse functions. In M. deBoer & D. Leontjev (Eds.), *Assessment and learning in content and language integrated learning (CLIL) classrooms* (pp. 229–251). Springer.
- deBoer, M., & Leontjev, D. (Eds.). (2020). *Assessment and learning in content and language integrated learning (CLIL) classrooms*. Springer. <https://doi.org/10.1007/978-3-030-54128-6>
- del Pozo, E. (in progress). *The learning of history contents in English in bilingual and non-bilingual settings in the Madrid region: Programmes and challenges* [PhD thesis]. Universidad Autónoma de Madrid, Madrid, Spain.
- del Pozo, E. (2019). CLIL in secondary classrooms: History contents on the move. In K. Tsuchiya & M. D. Pérez Murillo (Eds.), *Content and language integrated learning in Spanish and Japanese contexts* (pp. 125–151). Springer. [https://doi.org/10.1007/978-3-030-27443-6\\_6](https://doi.org/10.1007/978-3-030-27443-6_6)
- del Pozo, E., & Llinares, A. (2021). Assessing students' learning of history content in Spanish CLIL programmes: A content and language integrated perspective. In C. Hemmi & D. L. Banegas (Eds.), *International Perspectives on CLIL* (pp. 43–61). Springer.  
[https://doi.org/10.1007/978-3-030-70095-9\\_3](https://doi.org/10.1007/978-3-030-70095-9_3)

- Denman, J., van Schooten, E., & de Graaff, R. (2018). CLIL and bilingual education in the Netherlands. *Dutch Journal of Applied Linguistics*, 7(2), 203–226.  
<https://doi.org/10.1075/dujal.18005.den>
- Deschner, A., Eisele, N., Alavi, B., Demantowsky, M., Kenkmann, A., Popp, S., & Sauer, M. (2010). Bilingualer Unterricht im Sachfach Geschichte [bilingual teaching in the content subject history]: Ansatz und erste Ergebnisse eines Forschungsprojekts zur Lehrerkompetenz „Kompetenzorientierte Materialentwicklung“ [approach and first results of the research project "competency-based material development"]. *Zeitschrift Für Geschichtsdidaktik*, 9, 98–109.
- The Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8.
- Devos, N. J. (2015). *Peer interactions in new content and language integrated settings. Educational linguistics: Vol. 24*. Springer.
- Dewaele, J.-M. (2017). Why the dichotomy ‘L1 versus LX user’ is better than ‘native versus non-native speaker’. *Applied Linguistics*, 39(2), 236–240.  
<https://doi.org/10.1093/APPLIN/AMW055>
- Dijkstra, E. M., Walraven, A., Mooij, T., & Kirschner, P. A. (2017). Factors affecting intervention fidelity of differentiated instruction in kindergarten. *Research Papers in Education*, 32(2), 151–169. <https://doi.org/10.1080/02671522.2016.1158856>
- Dios Martínez-Agudo, J. de (2019). To what extent can CLIL learners' oral competence outcomes be explained by contextual differences? Updated empirical evidence from Spain. *Southern African Linguistics and Applied Language Studies*, 37(1), 27–40.  
<https://doi.org/10.2989/16073614.2019.1598878>
- Doiz, A., & Lasagabaster, D. (2021). An analysis of the use of cognitive discourse functions in English-medium history teaching at university. *English for Specific Purposes*, 62, 58–69.  
<https://doi.org/10.1016/j.esp.2020.12.002>
- Doiz, A., Lasagabaster, D., & Sierra, J. M. (2014). CLIL and motivation: The effect of individual and contextual variables. *The Language Learning Journal*, 42(2), 209–224.  
<https://doi.org/10.1080/09571736.2014.889508>
- Donato, R. (2016). Sociocultural theory and content-based foreign language instruction: Theoretical insights on the challenge of integration. In L. Cammarata (Ed.), *Content-based foreign language teaching: Curriculum and pedagogy for developing advanced thinking and literacy skills* (pp. 25–50). Routledge.
- Döring, V. (2020). Student voices on CLIL: Suggestions for improving compulsory CLIL education in Austrian technical colleges (HTL). *CELT Matters*, 4, 1–11.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford applied linguistics. Oxford University Press.
- The Douglas Fir Group (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal*, 100(S1), 19–47. <https://doi.org/10.1111/modl.12301>
- Dresing, T., & Pehl, T. (2015). *Praxisbuch Interview, Transkription & Analyse [practice book interview, transcription, and analysis]: Anleitungen und Regelsysteme für qualitative Forschende [manuals and regulatory systems for qualitative research]* (6th ed.). Eigenverlag.



- Duff, P. A., & Byrnes, H. (2019). SLA across disciplinary borders: Introduction to the special issue. *The Modern Language Journal*, 103, 3–5. <https://doi.org/10.1111/modl.12537>
- Dzik, D. (2020). Intercomprehension: A mere dream or a new way of learning in globalised world? *Politeja*, 16(3), 155–166. <https://doi.org/10.12797/Politeja.16.2019.60.10>
- ECML. (2020). *A pluriliteracies approach to teaching for learning*. European Centre for Modern Languages of the Council of Europe. <https://pluriliteracies.ecml.at/>
- Edelson, D. (2006). Balancing innovation and risk: Assessing design research proposals. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 100–106). Taylor & Francis Ltd. <http://public.eblib.com/choice/publicfullrecord.aspx?p=274476>
- Wissenschaftliche Erhebungen an Schulen [research at schools] (2017). <http://erlaesse.ssr-wien.gv.at/Erlaesse/tabid/58/Default.aspx?EntryId=8250>
- Ehlich, K., & Rehbein, J. (1986). *Muster und Institution [patterns and institution]: Untersuchungen zur schulischen Kommunikation [examining communication at school]. Kommunikation und Institution: Vol. 15*. Narr Francke Attempto.
- Eijkelhof, H. (2017, July 18). *An introduction to design-based educational research*. University of Vienna. Summerschool des Zentrums für LehrerInnenbildung, Spital am Phyrn. <https://lehrerinnenbildung.univie.ac.at/forschung/summer-school/summer-school-2017/>
- ELINET. (2015). *Teacher learning for European literacy education (TeL4ELE)*. European Commission. <http://www.eli-net.eu/good-practice/examples-of-good-practice/detail/project/teacher-learning-for-european-literacy-education-tel4ele>
- Ellin, J. (2004). Classification and categorization: A difference that makes a difference. *Library Trends*, 52(3), 515–540.
- Ethics Committee of the University of Vienna. (2021). *Mission statement*. <https://ethikkommission.univie.ac.at/en/mission-statement/>
- Euler, D. (2014). Design research: A paradigm under development. In D. Euler & P. F. E. Sloane (Eds.), *Design-based research* (pp. 15–43). Franz Steiner Verlag.
- Euler, D., & Sloane, P. F. E. (Eds.). (2014). *Design-based research*. Franz Steiner Verlag.
- European Commission. (1995). *White paper on education and training: Teaching and learning: Towards the learning society*. [http://europa.eu/documents/comm/white\\_papers/pdf/com95\\_590\\_en.pdf](http://europa.eu/documents/comm/white_papers/pdf/com95_590_en.pdf)
- Eurydice. (2006). *Content and language integrated learning (CLIL) at school in Europe*. European Commission.
- Eurydice. (2017). *Key data on teaching languages at school in Europe*. Brussels. European Commission.
- Evnitskaya, N. (2019). Constructing cognitive discourse functions in secondary CLIL classrooms in Spain. In K. Tsuchiya & M. D. Pérez Murillo (Eds.), *Content and language integrated learning in Spanish and Japanese contexts* (pp. 237–262). Springer. [https://doi.org/10.1007/978-3-030-27443-6\\_10](https://doi.org/10.1007/978-3-030-27443-6_10)
- Evnitskaya, N., & Dalton-Puffer, C. (2020). Cognitive discourse functions in CLIL classrooms: Eliciting and analysing students' oral categorizations in science and history. *International*

- Journal of Bilingual Education and Bilingualism*, 1–20.  
<https://doi.org/10.1080/13670050.2020.1804824>
- Evnitskaya, N., & Morton, T. (2011). Knowledge construction, meaning-making and interaction in CLIL science classroom communities of practice. *Language and Education*, 25(2), 109–127. <https://doi.org/10.1080/09500782.2010.547199>
- Fernández Fontecha, A. (2014). Receptive vocabulary knowledge and motivation in CLIL and EFL. *Revista De Lingüística Y Lenguas Aplicadas*, 9, 23–32.
- Fernández-Fontecha, A. (2015). Motivation and vocabulary breadth in CLIL and EFL contexts: Different age, same time of exposure. *Complutense Journal of English Studies*, 23, 79–96. [https://doi.org/10.5209/rev\\_CJES.2015.v23.51214](https://doi.org/10.5209/rev_CJES.2015.v23.51214)
- Fernández-Sanjurjo, J., Fernández-Costales, A., & Arias Blanco, J. M. (2017). Analysing students' content-learning in science in CLIL vs. non-CLIL programmes: Empirical evidence from Spain. *International Journal of Bilingual Education and Bilingualism*, 1(1), 1–14. <https://doi.org/10.1080/13670050.2017.1294142>
- Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.
- Filice, S. (2021). CLIL in pharmacology: Enabling student voice. *Latin American Journal of Content and Language Integrated Learning*, 13(2), 313–338. <https://doi.org/10.5294/laclil.2020.13.2.7>
- Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, 22(4), 16–23. <https://doi.org/10.3102/0013189X022004016>
- Flutter, J., & Rudduck, J. (2004). *Consulting pupils: What's in it for schools?* Routledge Falmer. <http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10101275>
- Fodor, J. A. (1995). *The elm and the expert: Mentalese and its semantics* (1st ed.). A Bradford book. MIT Press.
- Fraefel, U. (2014). *Professionalization of pre-service teachers through university-school partnerships*. University of Edinburgh. WERA Focal Meeting, Edinburgh. <https://doi.org/10.13140/rg.2.1.1979.5925>
- Freudenthal, H. (1991). *Revisiting mathematics education: China lectures*. Mathematics education library. Kluwer Academic Publishers.
- Fung, D., & Yip, V. (2014). The effects of the medium of instruction in certificate-level physics on achievement and motivation to learn. *Journal of Research in Science Teaching*, 51(10), 1219–1245. <https://doi.org/10.1002/tea.21174>
- Gablasova, D. (2014). Issues in the assessment of bilingually educated students: Expressing subject knowledge through L1 and L2. *Language Learning Journal*, 42(2), 151–164. <https://doi.org/10.1080/09571736.2014.891396>
- Gallardo del Puerto, F., & Gómez Lacabex, E. (2017). Oral production outcomes in CLIL: An attempt to manage amount of exposure. *European Journal of Applied Linguistics*, 5(1), 31–54. <https://doi.org/10.1515/eujal-2015-0035>
- Gautschi, P. (2015). *Guter Geschichtsunterricht [good history teaching]: Grundlagen, Erkenntnisse, Hinweise [foundations, insights, evidence]* (3rd rev. ed.). Geschichtsunterricht erforschen. Wochenschau Verlag.

- Geertz, C. (1993). *The interpretation of cultures*. Fontana Press.
- Gené-Gil, M., Juan-Garau, M., & Salazar-Noguera, J. (2015). Development of EFL writing over three years in secondary education: CLIL and non-CLIL settings. *Language Learning Journal*, 43(3), 286–303. <https://doi.org/10.1080/09571736.2015.1053278>
- Genesee, F. (2006). What do we know about bilingual education for majority-language students? In T. K. Bhatia & W. C. Ritchie (Eds.), *The handbook of bilingualism* (pp. 547–576). Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470756997.ch21>
- Genesee, F. (2013). Insights into bilingual education from research on immersion programs in Canada. In C. Abello-Contesse, P. M. Chandler, M. D. López-Jiménez, & R. Chacón-Beltrán (Eds.), *Bilingual and multilingual education in the 21st century* (pp. 24–41). Multilingual Matters.
- Gerns, P. (in progress). *Cognitive discourse functions in content and language integrated learning: An empirical study in secondary school science* [PhD thesis]. Universidad de Navarra.
- Gierlinger, E. M. (2015). ‘You can speak German, sir’: On the complexity of teachers' L1 use in CLIL. *Language and Education*, 29(4), 347–368. <https://doi.org/10.1080/09500782.2015.1023733>
- Gierlinger, E. M. (2021). L2 confidence in CLIL teaching: A tale of two teachers. In K. R. Talbot, M.-T. Gruber, & R. Nishida (Eds.), *The psychological experience of integrating content and language* (pp. 195–212). Multilingual Matters.
- Gierlinger, E. M., & Wagner, T. A. (2016). The more the merrier: Revisiting CLIL-based vocabulary growth in secondary education. *Latin American Journal of Content and Language Integrated Learning*, 9(1), 37–63. <https://doi.org/10.5294/laclil.2016.9.1.3>
- Gläser, R. (1990). *Fachtextsorten im Englischen [technical genres in English]* (Vol. 13). Narr Francke Attempto.
- Gómez Lacabex, E., & Gallardo-del-Puerto, F. (2020). Explicit phonetic instruction vs. implicit attention to native exposure: Phonological awareness of English schwa in CLIL. *International Review of Applied Linguistics in Language Teaching*, 58(4), 419–442.
- Goris, J., Denessen, E [Eddie], & Verhoeven, L. (2013). Effects of the content and language integrated learning approach to EFL teaching: A comparative study. *Written Language & Literacy*, 16(2), 186–207. <https://doi.org/10.1075/wll.16.2.03gor>
- Goris, J., Denessen, E [Eddie], & Verhoeven, L. (2017). The contribution of CLIL to learners’ international orientation and EFL confidence. *The Language Learning Journal*, 1–11. <https://doi.org/10.1080/09571736.2016.1275034>
- Goris, J., Denessen, E [EJPG], & Verhoeven, L. T. (2019). Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies. *European Educational Research Journal*, 18(6), 675–698. <https://doi.org/10.1177/1474904119872426>
- Graham, K., Choi, Y., Davoodi, A., Razmeh, S., & Dixon, L. (2018). Language and content outcomes of CLIL and EMI: A systematic review. *Latin American Journal of Content and Language Integrated Learning*, 11(1), 19–37.
- Granados, A., Lorenzo-Espejo, A., & Lorenzo, F. (2021). Evidence for the interdependence hypothesis: A longitudinal study of biliteracy development in a CLIL/bilingual setting.

- International Journal of Bilingual Education and Bilingualism*, 1–17.  
<https://doi.org/10.1080/13670050.2021.2001428>
- Gravemeijer, K., & Cobb, P. (2006). Design research from a learning design perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 17–51). Taylor & Francis Ltd.
- Groundwater-Smith, S., & Mockler, N. (2016). From data source to co-researchers? Tracing the shift from 'student voice' to student–teacher partnerships in educational action research. *Educational Action Research*, 24(2), 159–176.  
<https://doi.org/10.1080/09650792.2015.1053507>
- Gruber, M.-T., Lämmerer, A., Hofstadler, N., & Mercer, S. (2020). Flourishing or floundering? Factors contributing to CLIL primary teachers' wellbeing in Austria. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, 3(1), 19.  
<https://doi.org/10.5565/rev/clil.24>
- Gupta, K. C.-L. (2020). Researcher-teacher collaboration in adopting critical content and language integrated learning (CLIL): Processes, challenges, and outcomes. *Trabalhos Em Linguística Aplicada*, 59(1), 42–77. <https://doi.org/10.1590/010318136014125912020>
- Hahn, S. (2019). The problem with materials in CLIL: Needs and perspectives of Austrian CLIL history teachers. *CELT Matters*, 3, 17–24.  
[https://anglistik.univie.ac.at/fileadmin/user\\_upload/i\\_anglistik/Department/CELT/CELT\\_Matters/Hahn\\_3\\_2019\\_03.pdf](https://anglistik.univie.ac.at/fileadmin/user_upload/i_anglistik/Department/CELT/CELT_Matters/Hahn_3_2019_03.pdf)
- Halliday, M. A. K. (1975). *Learning how to mean: Explorations in the development of language*. Elsevier.
- Halliday, M. A. K. (1993). Towards a language-based theory of learning. *Linguistics and Education*, 5(2), 93–116. [https://doi.org/10.1016/0898-5898\(93\)90026-7](https://doi.org/10.1016/0898-5898(93)90026-7)
- Halliday, M. A. K., & Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective* (3rd ed.). *Social semiotic*. Oxford University Press.
- Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's introduction to functional grammar* (4th rev. ed.). Routledge.
- Hamidavi, N., Amiz, M. S., & Gorjian, B. (2016). The effect of the CLIL method on teaching reading comprehension to junior high school students. *Bulletin De La Société Royale Des Sciences De Liège*, 85, 1642–1652. <https://doi.org/10.25518/0037-9565.6361>
- Harmer, J. (2015). *The practice of English language teaching* (5th ed.). Pearson Longman.
- Hasberg, W., & Körber, A. (2003). Geschichtsbewusstsein dynamisch [dynamic historical consciousness]. In A. Körber (Ed.), *Geschichte - Leben - Lernen [history - living - learning]: Bodo von Borries zum 60. Geburtstag [to the 60th birthday of Bodo von Borries]* (pp. 179–202). Wochenschau Verlag.
- Hasenberger, T. (in progress). *English for the natural sciences: Developing and implementing a curriculum for a new subject at upper-secondary schools* [PhD thesis]. University of Vienna.
- Hasenberger, T. (2018). "The science of it ...": Designing and teaching a CLIL curriculum. *CELT Matters*, 2, 11–18. <https://anglistik.univie.ac.at/staff/teams-and-research-groups/eduling/celt-matters/contributions/>

- Heil, W. (2012). *Kompetenzorientierter Geschichtsunterricht [competency-based history education]* (2nd ed.). Kohlhammer.
- Heimes, A. (2011). *Psycholinguistic thought meets sociocultural theory: Die integrativen Zusammenhänge von Fachmethodik und Fremdsprachenlernen im bilingualen (Geschichts-)Unterricht [the integrative relations between subject didactics and learning foreign languages in the bilingual (history) classroom]*. *Inquiries in language learning*. Peter Lang.
- Heine, L. (2010). Fremdsprache und konzeptuelle Repräsentation[foreign language and conceptual representation]: Bilingualer Unterricht aus kognitiver Perspektive [bilingual education from a cognitive perspective]. In S. Dorff (Ed.), *Narr-Studienbücher* (pp. 199–212). Narr Francke Attempto.
- Heras, A., & Lasabaster, D. (2015). The impact of CLIL on affective factors and vocabulary learning. *Language Teaching Research*, 19(1), 70–88.  
<https://doi.org/10.1177/1362168814541736>
- Hermanns, H. (2010). Interviewing as an activity. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 209–213). SAGE Publications.
- Hobsbawm, E. (1997). *The age of extremes: The short twentieth century, 1914-1991*. Abacus.
- Hoffmann, L. (1988). *Vom Fachwort zum Fachtext [from technical terminology to technical texts]: Beiträge zur angewandten Linguistik [contributions to applied linguistics]*. *Forum für Fachsprachen-Forschung: Bd. 5*. Narr Francke Attempto.
- Holzman, L. (2018). Zones of proximal development: Mundane and magical. In J. P. Lantolf, M. E. Poehner, & m. Swain (Eds.), *Routledge handbooks in applied linguistics. The Routledge handbook of sociocultural theory and second language development* (1st ed., pp. 42–55). Routledge an imprint of Taylor and Francis.
- Hu, J., & Gao, X. (2020). Understanding subject teachers' language-related pedagogical practices in content and language integrated learning classrooms. *Language Awareness*, 6(2), 1–20.  
<https://doi.org/10.1080/09658416.2020.1768265>
- Hughes, S. P., & Madrid, D. (2020). The effects of CLIL on content knowledge in monolingual contexts. *The Language Learning Journal*, 48(1), 48–59.  
<https://doi.org/10.1080/09571736.2019.1671483>
- Hunt, M. (2011). UK teachers' and learners' experiences of CLIL resulting from the EU-funded project ECLILT. *Latin American Journal of Content and Language Integrated Learning*, 4(1), 27–39. <https://doi.org/10.5294/lacil.2011.4.1.3>
- Hüttner, J., Dalton-Puffer, C., & Smit, U. (2013). The power of beliefs: Lay theories and their influence on the implementation of CLIL programmes. *International Journal of Bilingual Education and Bilingualism*, 16(3), 267–284.  
<https://doi.org/10.1080/13670050.2013.777385>
- Hüttner, J., & Rieder-Bünemann, A. (2010). A cross-sectional analysis of oral narratives by children with CLIL and non-CLIL instruction. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, pp. 61–80). John Benjamins.
- Hüttner, J., & Smit, U. (2018). Negotiating political positions: Subject-specific oral language use in CLIL classrooms. *International Journal of Bilingual Education and Bilingualism*, 21(3), 287–302. <https://doi.org/10.1080/13670050.2017.1386616>

- IBM Corp. (2019). *SPSS statistics for Windows* (Version 26) [Computer software]. Armonk.  
<https://www.ibm.com/at-de/products/spss-statistics>
- Iglesias Diéguez, K., & Martínez-Adrián, M. (2017). The influence of CLIL on receptive vocabulary: A preliminary study. *Journal of English Studies*, 15, 107.  
<https://doi.org/10.18172/jes.3210>
- The Jamovi Project. (2021). *Jamovi* (Version 1.8) [Computer software]. <https://www.jamovi.org>
- Jäppinen, A.-K. (2005). Thinking and content learning of mathematics and science as cognitional development in content and language integrated learning (CLIL): Teaching through a foreign language in Finland. *Language and Education*, 19(2), 147–168.  
<https://doi.org/10.1080/09500780508668671>
- Järvinen, H.-M. (2010). Language as a meaning making resource in learning and teaching content. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, pp. 145–168). John Benjamins.
- Jeismann, K.-E. (1980). Geschichtsbewußtsein [historical consciousness]: Überlegungen zur zentralen Kategorie eines neuen Ansatzes der Geschichtsdidaktik [reflections on the central category of a new approach to history didactics]. In H. Süßmuth (Ed.), *Uni-Taschenbücher. Geschichtsdidaktische Positionen [positions in history didactics]: Bestandsaufnahme und Neuorientierung [taking stock and moving forward]* (Vol. 954, pp. 179–222). F. Schöningh.
- Jexenflicker, S., & Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, pp. 169–190). John Benjamins.
- Jiménez Catalán, R. M [Rosa M.], & Agustín-Llach, M. P. (2017). CLIL or time? Lexical profiles of CLIL and non-CLIL EFL learners. *System*, 66, 87–99.  
<https://doi.org/10.1016/j.system.2017.03.016>
- Johnson, R. B., & Christensen, L. (2012). *Educational research: Quantitative, qualitative, and mixed approaches* (4th ed.). SAGE Publications.
- Juan-Garau, M., Prieto-Arranz, J. I., & Salazar-Noguera, J. (2015). Lexico-grammatical development in secondary education CLIL learners. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Educational linguistics. Content-based language learning in multilingual educational environments* (Vol. 23, pp. 179–195). Springer.  
[https://doi.org/10.1007/978-3-319-11496-5\\_11](https://doi.org/10.1007/978-3-319-11496-5_11)
- Jung, E. (2010). *Kompetenzerwerb [acquisition of competences]: Grundlagen, Didaktik, Überprüfbarkeit [basics, didactics, testability]*. Oldenbourg Verlag.
- Kelly, A. (2006). Quality criteria for design research: Evidence and commitments. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 107–118). Taylor & Francis Ltd.  
<http://public.eblib.com/choice/publicfullrecord.aspx?p=274476>
- Kidd, R. (1996). Teaching academic language functions at the secondary level. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, 52(2), 285–305.
- Kilbey, E., Ward, C., & Ting, T. (2018). *Talent*. Cambridge University Press.

- Kim, Y. K., Hutchison, L. A., & Winsler, A. (2015). Bilingual education in the United States: An historical overview and examination of two-way immersion. *Educational Review*, 67(2), 236–252. <https://doi.org/10.1080/00131911.2013.865593>
- Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen [competency-models to apprehend individual learning outcomes and to balance educational processes]. *Zeitschrift Für Pädagogik*, 52, 876–903.
- Knorr, P., & Schramm, K. (2016). Triangulation. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik [research methods in language teaching research]: Ein Handbuch [a handbook]* (1st ed., pp. 90–97). Narr Francke Attempto.
- Kölbl, C., & Konrad, L. (2015). Historical consciousness in Germany: Concept, implementation, assessment. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 17–28). Routledge.
- Kong, S., Hoare, P., & Chi, Y. (2011). Immersion education in China: Teachers' perspectives. *Frontiers of Education in China*, 6(1), 68–91. <https://doi.org/10.1007/s11516-011-0122-6>
- Koopman, G. J., Skeet, J., & de Graaff, R. (2014). Exploring content teachers' knowledge of language pedagogy: A report on a small-scale research project in a Dutch CLIL context. *The Language Learning Journal*, 42(2), 123–136. <https://doi.org/10.1080/09571736.2014.889974>
- Körber, A. (2007a). Graduierung [graduation]: Die Unterscheidung von Niveaus der Kompetenzen Historischen Denkens [differentiating levels of historical competences]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 415–472). ars una.
- Körber, A. (2007b). Grundbegriffe und Konzepte [fundamental terms and concepts]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 54–86). ars una.
- Körber, A., & Meyer-Hamme, J. (2015). Historical thinking, competencies, and their measurement: Challenges and approaches. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 89–101). Routledge.
- Körber, A., Schreiber, W., & Schöner, A. (Eds.). (2007). *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]*. ars una.
- KPH Vienna/ Krems. (2020). *Fortbildungsprogramm NÖ 2020/2021 [in-service teacher training programme lower Austria 2020/2021]*.
- Krashen, S. D. (1987). *Principles and practice in second language acquisition /Stephen D. Krashen. Language teaching methodology series*. Prentice-Hall Internat.

- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory and Practice*, 41(4), 212–218.
- Kuckartz, U. (2016). *Qualitative Inhaltsanalyse [qualitative text analysis]: Methoden, Praxis, Computerunterstützung [methods, practice, and using software]* (3rd ed.). *Grundlagentexte Methoden*. Beltz Juventa.
- Kühberger, C. (2011). Operatoren als strukturierende Elemente von Aufgabenstellungen für Geschichte, Sozialkunde und Politische Bildung [performative verbs as structuring elements for tasks in history, social studies, and political education]. In *Die kompetenzorientierte Reifeprüfung Geschichte und Sozialkunde, Politische Bildung [the competency-based final exam in history, social studies, and political education]* (pp. 15–20). Bundesministerium für Bildung und Frauen.  
[https://www.bmb.gv.at/schulen/unterricht/ba/reifepruefung\\_ahs\\_lfgsk\\_21067.pdf](https://www.bmb.gv.at/schulen/unterricht/ba/reifepruefung_ahs_lfgsk_21067.pdf)
- Kühberger, C. (2015). *Kompetenzorientiertes historisches und politisches Lernen [competency-based historical and political learning]: Methodische und didaktische Annäherungen für Geschichte, Sozialkunde und Politische Bildung [methodological and didactical approaches for history, social studies, and political education]* (3rd rev. ed.). *Österreichische Beiträge zur Geschichtsdidaktik: Band 2*. StudienVerlag.
- Kühberger, C. (2019). History education research in Austria. In M. Köster, H. Thünemann, & M. Zülsdorf-Kersting (Eds.), *Researching history education: International perspectives and disciplinary traditions* (2nd ed., pp. 153–177). Wochenschau Verlag.
- Kühberger, C., & Windischbauer, E. (2012). Kommentar zum Lehrplan der AHS-Unterstufe und Hauptschule [comment on the curricula for lower secondary education (AHS and HS)]: Geschichte und Sozialkunde/Politische Bildung [history, social studies, and political education]. In B. Dmytrasz, A. Ecker, I. Ecker, & F. Öhl (Eds.), *Fachdidaktik Geschichte, Sozialkunde und Politische Bildung [subject didactics history, social studies, and political education]: Modelle, Texte, Beispiele [models, texts, and examples]* (pp. 6–16). Austrian Federal Ministry of Education.
- Lackner, M. (2012). *The use of subject-related discourse functions in upper secondary CLIL history classes* [diploma thesis]. University of Vienna, Vienna.
- Lahuerta, A. (2015). The written competence of Spanish secondary education students in bilingual and non-bilingual programs. *Porta Linguarum*, 24, 47–61.
- Lahuerta, A. (2020). Analysis of accuracy in the writing of EFL students enrolled on CLIL and non-CLIL programmes: The impact of grade and gender. *The Language Learning Journal*, 48(2), 121–132. <https://doi.org/10.1080/09571736.2017.1303745>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863.  
<https://doi.org/10.3389/fpsyg.2013.00863>
- Lamsfuß-Schenk, S. (2010). Inhalt und Sprache [content and language]: Vom Einfluss des Fremdsprachengebrauchs auf das Lernen im Sachfach [of the impact of foreign language use on the learning in content subjects]. In S. Dorff (Ed.), *Bilingualer Sachfachunterricht in der Sekundarstufe [bilingual content teaching in secondary education]: Eine Einführung [an introduction]* (pp. 213–227). Narr Francke Attempto.



- Lantolf, J. P., Poehner, M. E., & Swain, m. (Eds.). (2018). *Routledge handbooks in applied linguistics. The Routledge handbook of sociocultural theory and second language development* (1st ed.). Routledge an imprint of Taylor and Francis.  
<https://www.routledgehandbooks.com/doi/10.4324/9781315624747>  
<https://doi.org/10.4324/9781315624747>
- Lasabaster, D. (2013). The use of the L1 in CLIL classes: The teachers' perspective. *Latin American Journal of Content and Language Integrated Learning*, 6(2), 1–21.  
<https://doi.org/10.5294/laclil.2013.6.2.1>
- Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal*, 1(1), 30–41.  
<https://doi.org/10.2174/1874913500801010030>
- Lasagabaster, D. (2011). English achievement and student motivation in CLIL and EFL settings. *Innovation in Language Learning and Teaching*, 5(1), 3–18.  
<https://doi.org/10.1080/17501229.2010.519030>
- Lasagabaster, D., & Doiz, A. (2015). A longitudinal study on the impact of CLIL on affective factors. *Applied Linguistics*, 1(4), amv059. <https://doi.org/10.1093/applin/amv059>
- Lasagabaster, D., Doiz, A., Gómez Lacabex, E., & Kopinska, M. (2021). *Learning history in English: Language-related materials for students*. Servicio Editorial UPV/EHU.
- Lasagabaster, D., & Sierra, J. M. (2009). Language attitudes in CLIL and traditional EFL classes. *International CLIL Research Journal*, 1(2), 4–17. <http://www.icrj.eu/12-73>
- Lasagabaster, D., & Sierra, J. M. (2010). Immersion and CLIL in English: More differences than similarities. *ELT Journal*, 64(4), 367–375. <https://doi.org/10.1093/elt/ccp082>
- Lee, P. (2017). History education and historical literacy. In I. Davies (Ed.), *Debates in history teaching* (55–65). Routledge.
- Lemke, J. L. (1990). *Talking science: language, learning, and values. Language and educational processes*. Ablex Pub. Corp.
- Lialikhova, D. (2019). “We can do it together!” – But can they? How Norwegian ninth graders co-constructed content and language knowledge through peer interaction in CLIL. *Linguistics and Education*, 54, 1–19. <https://doi.org/10.1016/j.linged.2019.100764>
- Llinares, A. G., & Morton, T. (2010). Historical explanations as situated practice in content and language integrated learning. *Classroom Discourse*, 1(1), 46–65.  
<https://doi.org/10.1080/19463011003750681>
- Llinares, A. G., & Morton, T. (2017). Speech function analysis to explore CLIL students’ spoken language for knowledge construction. In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 125–144). John Benjamins.
- Llinares, A. G., Morton, T., & Whittaker, R. (2012). *The roles of language in CLIL. Cambridge language teaching library*. Cambridge University Press.
- Llinares, A. G., & Nikula, T. (2016). Teacher and student evaluative language in CLIL across contexts: Integrating SFL and pragmatic approaches. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (pp. 189–210). Multilingual Matters.

- Llinares, A. G., & Pascual Peña, I. (2015). A genre approach to the effect of academic questions on CLIL students' language production. *Language and Education*, 29(1), 15–30.  
<https://doi.org/10.1080/09500782.2014.924964>
- Llinares, A. G., & Whittaker, R. (2010). Writing and speaking in the history class: A comparative analysis of CLIL and first language contexts. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, 125–124). John Benjamins.  
<https://benjamins.com/catalog/aals.7.07lli>
- Lo, Y. Y., & Jeong, H. (2018). Impact of genre-based pedagogy on students' academic literacy development in content and language integrated learning (CLIL). *Linguistics and Education*, 47, 36–46. <https://doi.org/10.1016/j.linged.2018.08.001>
- Lo, Y. Y., Lin, A. M. Y., & Liu, Y. (2020). Exploring content and language co-construction in CLIL with semantic waves. *International Journal of Bilingual Education and Bilingualism*, 5(1), 1–22. <https://doi.org/10.1080/13670050.2020.1810203>
- Lopez Ornat, S. (2012). Language acquisition and development. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1718–1721). Springer US. [https://doi.org/10.1007/978-1-4419-1428-6\\_298](https://doi.org/10.1007/978-1-4419-1428-6_298)
- López-Medina, B. (2016). Developing a CLIL textbook evaluation checklist. *Latin American Journal of Content and Language Integrated Learning*, 9(1), 159–173.  
<https://doi.org/10.5294/laclil.2016.9.1.7>
- Lorenzo, F. (2017). Historical literacy in bilingual settings: Cognitive academic language in CLIL history narratives. *Linguistics and Education*, 37, 32–41.  
<https://doi.org/10.1016/j.linged.2016.11.002>
- Lorenzo, F., & Dalton-Puffer, C. (2016). Historical literacy in CLIL: Telling the past in a second language. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (pp. 55–74). Multilingual Matters.
- Lorenzo, F., Granados, A., & Ávila, I. (2019). The development of cognitive academic language proficiency in multilingual education: Evidence of a longitudinal study on the language of history. *Journal of English for Academic Purposes*, 41, 100767.  
<https://doi.org/10.1016/j.jeap.2019.06.010>
- Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach. Language learning and language teaching: Vol. 18*. John Benjamins.
- Lyster, R. (2015). Using form-focused tasks to integrate language across the immersion curriculum. *System*, 54, 4–13. <https://doi.org/10.1016/j.system.2014.09.022>
- Lyster, R. (2017). Introduction to part I: SLA perspectives on learning and teaching language through content. In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 19–32). John Benjamins.
- Macnaught, L., Maton, K., Martin, J. R., & Matruglio, E. (2013). Jointly constructing semantic waves: Implications for teacher training. *Linguistics and Education*, 24(1), 50–63.  
<https://doi.org/10.1016/j.linged.2012.11.008>

- Madrid, D., & Pérez Cañado, M. L. (2018). Innovations and challenges in attending to diversity through CLIL. *Theory into Practice*, 57(3), 241–249. <https://doi.org/10.1080/00405841.2018.1492237>
- Mahan, K. R. (2020). The comprehending teacher: Scaffolding in content and language integrated learning (CLIL). *The Language Learning Journal*, 18(4), 1–15. <https://doi.org/10.1080/09571736.2019.1705879>
- Mahan, K. R., Brevik, L. M., & Ødegaard, M. (2018). Characterizing CLIL teaching: New insights from a lower secondary classroom. *International Journal of Bilingual Education and Bilingualism*, 118(110306), 1–18. <https://doi.org/10.1080/13670050.2018.1472206>
- Maposa, M., & Wassermann, J. (2009). Conceptualizing historical literacy: A review of the literature. *Yesterday & Today*, 4(41-66).
- Marsh, D., & Langé, G. (Eds.). (2000). *Using languages to learn and learning to use languages: An introduction to content and language integrated learning for parents and young people*. Univ. of Jyväskylä.
- Martin, J. R., & Rose, D. (2008). *Genre relations: Mapping culture. Equinox textbooks and surveys in linguistics*. Equinox.
- Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan. <http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10262882>
- Martínez Adrián, M., & Gutiérrez Mangado, M. J. (2015). L1 use, lexical richness, accuracy and syntactic complexity in the oral production of CLIL and non-CLIL learners of English. *Atlantis*, 37(2), 175–197.
- Maset, M. (2015). *Bilingualer Geschichtsunterricht [bilingual history education]: Didaktik und Praxis [didactics and practice]* (1st ed.). *Geschichte im Unterricht [history in the classroom]*. Verlag W. Kohlhammer.
- Massler, U. (2012). Primary CLIL and its stakeholders: What children, parents and teachers think of the potential merits and pitfalls of CLIL modules in primary teaching. *International CLIL Research Journal*, 1(4), 36–46.
- Maton, K. (2013). Making semantic waves: A key to cumulative knowledge-building. *Linguistics and Education*, 24(1), 8–22. <https://doi.org/10.1016/j.linged.2012.11.005>
- Maxwell-Reid, C., & Lau, K. (2016). Genre and technicality in analogical explanations: Hong Kong's English language textbooks for junior secondary science. *Journal of English for Academic Purposes*, 23, 31–46. <https://doi.org/10.1016/j.jeap.2016.05.005>
- Mayring, P. (2015). *Qualitative Inhaltsanalyse [qualitative content analysis]: Grundlagen und Techniken [basics and techniques]* (12th rev. ed.). *Beltz Pädagogik*. Beltz.
- McCabe, A., & Whittaker, R. (2017). Genre and appraisal in CLIL history texts: Developing the voice of the historian. In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 105–124). John Benjamins.
- McCullough, D. (1984). *Graduation speech Wesleyan University*. Wesleyan University, Middletown. <https://www.nytimes.com/1984/06/04/nyregion/historian-addresses-wesleyan.html>
- McKenney, S., Nieveen, N., & van den Akker, J. (2006). Design research from a curriculum perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.),

- Educational design research* (pp. 67–90). Taylor & Francis Ltd.  
<http://public.eblib.com/choice/publicfullrecord.aspx?p=274476>
- McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. Routledge.
- McKenney, S., & Reeves, T. C. (2014). Methods of evaluation and reflection in design research. In D. Euler & P. F. E. Sloane (Eds.), *Design-based research* (pp. 141–155). Franz Steiner Verlag.
- Mearns, T. (2012). Using CLIL to enhance pupils' experience of learning and raise attainment in German and health education: A teacher research project. *The Language Learning Journal*, 40(2), 175–192. <https://doi.org/10.1080/09571736.2011.621212>
- Mearns, T., de Graaff, R., & Coyle, D. (2017). Motivation for or from bilingual education? A comparative study of learner views in the Netherlands. *International Journal of Bilingual Education and Bilingualism*, 20(4), 1–14.  
<https://doi.org/10.1080/13670050.2017.1405906>
- Megill, A. (1994). Jörn Rüsen's theory of historiography between modernism and rhetoric of inquiry. *History and Theory*, 33(1), 39–60.
- Mehisto, P., Marsh, D., & Frigols, M. J. (2009). *Uncovering CLIL: Content and language integrated learning in bilingual and multilingual education*. Macmillan books for teachers. Macmillan and Hueber.
- Méndez García, María del Carmen, & Pavon Vazquez, V. (2012). Investigating the coexistence of the mother tongue and the foreign language through teacher collaboration in CLIL contexts: Perceptions and practice of the teachers involved in the plurilingual programme in Andalusia. *International Journal of Bilingual Education and Bilingualism*, 15(5), 573–592. <https://doi.org/10.1080/13670050.2012.670195>
- Merisuo-Storm, T. (2007). Pupils' attitudes towards foreign-language learning and the development of literacy skills in bilingual education. *Teaching and Teacher Education*, 23(2), 226–235. <https://doi.org/10.1016/j.tate.2006.04.024>
- Mesquida, F., & Juan-Garau, M. (2013). CLIL instruction and its effects on the development of negotiation strategies. *A Journal of English and American Studies*, 47(1137-6368), 125–144.
- Mewald, C. (2007). A comparison of oral foreign language performance of learners in CLIL and in mainstream classes at lower secondary level in Lower Austria. In C. Dalton-Puffer (Ed.), *Sprache im Kontext. Empirical perspectives on CLIL classroom discourse* (Vol. 26, pp. 139–177). Peter Lang.
- Meyer, O. (2013). Introducing the CLIL-pyramid: Key strategies and principles for CLIL planning and teaching. In M. Eisenmann (Ed.), *Anglistische Forschungen: Bd. 420. Basic issues in EFL teaching and learning* (2nd ed., pp. 295–313). Winter.
- Meyer, O., & Coyle, D. (2017). Pluriliteracies teaching for learning: Conceptualizing progression for deeper learning in literacies development. *European Journal of Applied Linguistics*, 5(2), 199–222. <https://doi.org/10.1515/eujal-2017-0006>
- Meyer, O., Coyle, D., Halbach, A., Schuck, K., & Ting, T. (2015). A pluriliteracies approach to content and language integrated learning: Mapping learner progressions in knowledge construction and meaning-making. *Language, Culture and Curriculum*, 28(1), 41–57.  
<https://doi.org/10.1080/07908318.2014.1000924>

- Meyer-Hamme, J. (2007). Historische Kompetenzen empirisch [empirical evidence of historical competences]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 674–693). ars una.
- Meyerhöffer, N., & Dreesmann, D. C. (2019). English-bilingual biology for standard classes development, implementation and evaluation of an English-bilingual teaching unit in standard German high school classes. *International Journal of Science Education*, 41(10), 1366–1386. <https://doi.org/10.1080/09500693.2019.1607620>
- Milla, R., & García Mayo, M. d. P. (2021). Teachers' and learners' beliefs about corrective feedback compared with teachers' practices in CLIL and EFL. In K. R. Talbot, M.-T. Gruber, & R. Nishida (Eds.), *The psychological experience of integrating content and language* (pp. 112–132). Multilingual Matters. <https://doi.org/10.21832/9781788924306-012>
- Mitra, D. (2018). Student voice in secondary schools: The possibility for deeper change. *Journal of Educational Administration*, 56(5), 473–487. <https://doi.org/10.1108/JEA-01-2018-0007>
- Moate, J. (2010). The integrated nature of CLIL: A sociocultural perspective. *International CLIL Research Journal*, 1(3), 38–45.
- Moate, J. (2011). The impact of foreign language mediated teaching on teachers' sense of professional integrity in the CLIL classroom. *European Journal of Teacher Education*, 34(3), 333–346. <https://doi.org/10.1080/02619768.2011.585023>
- Moe, E., Härmälä, M., Pascoal, J., Meilutem Ramonienė, & Kristmanson, P. (2015). *Language skills for successful subject learning: CEFR-linked descriptors for mathematics and history/civics*. Council of Europe Publishing.
- Mohan, B. (1986). *Language and content. The Addison-Wesley second language professional library series*. Addison-Wesley.
- Mohan, B., Leung, C., & Slater, T. (2010). Assessing language and content: A functional perspective. In A. Paran & L. Sercu (Eds.), *New perspectives on language and education. Testing the untestable in language education* (1st ed., Vol. 17, pp. 217–240). Channel View Publications.
- Möller, V. (2017). *Language acquisition in CLIL and non-CLIL settings: Learner corpus and experimental evidence on passive constructions. Studies in corpus linguistics: Vol. 80*. John Benjamins.
- Moore, J., Schleppegrell, M. J., & Palincsar, A. S. (2018). Discovering disciplinary linguistic knowledge with English learners and their teachers: Applying systemic functional linguistics concepts through design-based research. *TESOL Quarterly*, 52(4), 1022–1049. <https://doi.org/10.1002/tesq.472>
- Moore, P. (2011). Collaborative interaction in turn-taking: A comparative study of European bilingual (CLIL) and mainstream (MS) foreign language learners in early secondary education. *International Journal of Bilingual Education and Bilingualism*, 14(5), 531–549. <https://doi.org/10.1080/13670050.2010.537741>

- Moore, P., & Lorenzo, F. (2007). Adapting authentic materials for CLIL classrooms: An empirical study. *VIEWS: Vienna English Working Papers*, 16(3), 28–35.  
[http://anglistik.univie.ac.at/fileadmin/user\\_upload/dep\\_anglist/weitere\\_Uploads/Views/Views\\_0703.pdf](http://anglistik.univie.ac.at/fileadmin/user_upload/dep_anglist/weitere_Uploads/Views/Views_0703.pdf)
- Moore, P., & Lorenzo, F. (2015). Task-based learning and content and language integrated learning materials design: Process and product. *The Language Learning Journal*, 43(3), 334–357. <https://doi.org/10.1080/09571736.2015.1053282>
- Morton, T. (2010). Using a genre-based approach to integrating content and language in CLIL. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, pp. 81–104). John Benjamins.
- Morton, T. (2013). Critically evaluating materials for CLIL: Practitioners' Practices and perspectives. In J. Gray (Ed.), *Critical perspectives on language teaching materials* (pp. 111–136). Palgrave Macmillan UK.  
[http://link.springer.com/10.1057/9781137384263\\_6](http://link.springer.com/10.1057/9781137384263_6)
- Morton, T. (2018). Reconceptualizing and describing teachers' knowledge of language for content and language integrated learning (CLIL). *International Journal of Bilingual Education and Bilingualism*, 21(3), 275–286.  
<https://doi.org/10.1080/13670050.2017.1383352>
- Morton, T. (2019). Teacher education in content-based language education. In S. Walsh & S. Mann (Eds.), *The Routledge handbook of English language teacher education* (pp. 169–183). Routledge.
- Morton, T. (2020). Cognitive discourse functions: A bridge between content, literacy and language for teaching and assessment in CLIL. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, 3(1), 7. <https://doi.org/10.5565/rev/clil.33>
- Morton, T., & Jakonen, T. (2016). Integration of language and content through languaging in CLIL classroom interaction: A conversation analysis perspective. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (pp. 171–188). Multilingual Matters.
- Morton, T., & Llinares, A. G. (2017). Content and language integrated learning (CLIL): Type of programme or pedagogical model? In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 1–18). John Benjamins.
- Morton, T., & Llinares, A. G. (2018). Students' use of evaluative language in L2 English to talk and write about history in a bilingual education programme. *International Journal of Bilingual Education and Bilingualism*, 21(4), 496–508.  
<https://doi.org/10.1080/13670050.2016.1192101>
- Nashaat Sobhy, N. (2017). Investigating Pragmatics in CLIL through students' requests. In A. G. Llinares & T. Morton (Eds.), *Language learning & language teaching. Applied linguistics perspectives on CLIL* (Vol. 47, pp. 67–90). John Benjamins.
- Nashaat Sobhy, N. (2018). Operationalizing “defining” from a cognitive discourse perspective for learners' use. In S. Anwaruddin (Ed.), *Knowledge mobilization in TESOL* (pp. 94–112). Brill. [https://doi.org/10.1163/9789004392472\\_007](https://doi.org/10.1163/9789004392472_007)

- Nashaat-Sobhy, N., & Llinares, A. (2020). CLIL students' definitions of historical terms. *International Journal of Bilingual Education and Bilingualism*, 9, 1–14. <https://doi.org/10.1080/13670050.2020.1798868>
- National Center for History in the Schools. (1996). *National standards for history: Basic edition*. UCLA. <https://phi.history.ucla.edu/nchs/history-standards/>
- Navarro Gil, N. (2019). The effects of a content-based language course on students' academic vocabulary production. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, 2(2), 25. <https://doi.org/10.5565/rev/clil.30>
- Nieto Moreno de Diezmas, E. (2018). The acquisition of L2 listening comprehension skills in primary and secondary education settings: A comparison between CLIL and non-CLIL student performance. *RLA. Revista De Lingüística Teórica Y Aplicada*, 56(2), 13–34. <https://doi.org/10.4067/S0718-48832018000200013>
- Nikula, T. (2010). Effects of CLIL on a teacher's classroom language use. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, pp. 105–124). John Benjamins.
- Nikula, T., Dafouz, E., Moore, P., & Smit, U. (Eds.). (2016). *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education*. Multilingual Matters.
- Nikula, T., Skinnari, K., & Mård-Miettinen, K. (accepted). Diversity in CLIL as experienced by Finnish CLIL teachers and students. *International Journal of Bilingual Education and Bilingualism*.
- O Ceallaigh, T., Ní Mhurchú, S., & Ní Chróinín, D. (2017). Balancing content and language in CLIL: The experiences of teachers and learners. *Journal of Immersion and Content-Based Language Education*, 5(1), 58–86. <https://doi.org/10.1075/jicb.5.1.03oce>
- Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism. A harvest book: Vol. 29*. Harcourt, Brace and World.
- Ohlberger, S., & Wegner, C. (2017). Motivational changes due to the implementation of a bilingual module in biology. *Journal of Innovation in Psychology, Education and Didactics*, 21(2), 149–176.
- Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8(4), 434–447. <https://doi.org/10.1037/1082-989X.8.4.434>
- Olsson, E. (2015). Progress in English academic vocabulary use in writing among CLIL and non-CLIL students in Sweden. *Moderna Språk*, 109(2), 51–74.
- Olsson, E., & Sylvén, L. (2015). Extramural English and academic vocabulary: A longitudinal study of CLIL and non-CLIL students in Sweden. *Apples - Journal of Applied Language Studies*, 9(2), 77–103. <https://doi.org/10.17011/apples/urn.201512234129>
- Otto, A., & Estrada, J. L. (2019). Towards an understanding of CLIL assessment practices in a European context: Main assessment tools and the role of language in content subjects. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, 2(1), 31. <https://doi.org/10.5565/rev/clil.11>

- Otwinowska, A., & Foryś, M. (2017). They learn the CLIL way, but do they like it? Affectivity and cognition in upper-primary CLIL classes. *International Journal of Bilingual Education and Bilingualism*, 20(5), 457–480. <https://doi.org/10.1080/13670050.2015.1051944>
- Ouazizi, K. (2016). The effects of CLIL education on the subject matter (mathematics) and the target language (English). *Latin American Journal of Content and Language Integrated Learning*, 9(1), 110–137. <https://doi.org/10.5294/laclil.2016.9.1.5>
- Oxbrow, G. L. (2018). Students' perspectives on CLIL programme development: A quantitative analysis. *Porta Linguarum*, 29, 137–158.
- Pandel, H.-J. (1987). Dimensionen des Geschichtsbewußtseins [dimensions of historical consciousness]: Ein Versuch, seine Struktur für Empirie und Pragmatik diskutierbar zu machen [an attempt at making its structure discussable for empiricism and pragmatics]. *Geschichtsdidaktik*, 12(130-142).
- Pandel, H.-J. (2012). *Geschichtsunterricht nach PISA [history education after PISA]: Kompetenzen, Bildungsstandards und Kerncurricula [competences, educational standards and core curricula]* (3rd ed.). Forum Historisches Lernen. Wochenschau Verlag.
- Pandel, H.-J. (2017). *Geschichtsdidaktik [history didactics]: Eine Theorie für die Praxis [a theory for practice]* (2nd ed.). Forum Historisches Lernen. Wochenschau Verlag.
- Paniagua, A., & Istance, D. (2018). *Teachers as designers of learning environments*. OECD. <https://doi.org/10.1787/9789264085374-en>
- Pappa, S., Moate, J., Ruohotie-Lyhty, M., & Eteläpelto, A. (2017a). Teacher agency within the Finnish CLIL context: Tensions and resources. *International Journal of Bilingual Education and Bilingualism*, 30, 1–21. <https://doi.org/10.1080/13670050.2017.1286292>
- Pappa, S., Moate, J., Ruohotie-Lyhty, M., & Eteläpelto, A. (2017b). Teachers' pedagogical and relational identity negotiation in the Finnish CLIL context. *Teaching and Teacher Education*, 65, 61–70. <https://doi.org/10.1016/j.tate.2017.03.008>
- Paran, A. (2013). Content and language integrated learning: Panacea or policy borrowing myth? *Applied Linguistics Review*, 4(2), 317–342. <https://doi.org/10.1515/applirev-2013-0014>
- Pastrana, A., Llinares, A. G., & Pascual, I. (2018). Students' language use for co-construction of knowledge in CLIL group-work activities: A comparison with L1 settings. *Zeitschrift Für Erziehungswissenschaft*, 21(1), 49–70. <https://doi.org/10.1007/s11618-017-0802-y>
- Pavón Vázquez, V. (2018). Innovations and challenges in CLIL research: Exploring the development of subject-specific literacies. *Theory into Practice*, 57(3), 204–211. <https://doi.org/10.1080/00405841.2018.1484035>
- Pavón Vázquez, V., & Méndez García, María del Carmen (2017). Analysing teachers' roles regarding cross-curricular coordination in content and language integrated learning (CLIL). *Journal of English Studies*, 15, 235. <https://doi.org/10.18172/jes.3227>
- Pecorari, D. (2018). English as a foreign Language (EFL) versus English as a second language (ESL) writing. In J. I. Lontas, T. International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching* (pp. 1–6). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118784235.eelt0507>
- Pérez Cañado, M. L. (submitted). Guest editorial. *International Journal of Bilingual Education and Bilingualism*.



- Pérez Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, 15(3), 315–341.  
<https://doi.org/10.1080/13670050.2011.630064>
- Pérez Cañado, M. L. (2016a). Are teachers ready for CLIL? Evidence from a European study. *European Journal of Teacher Education*, 39(2), 202–221.  
<https://doi.org/10.1080/02619768.2016.1138104>
- Pérez Cañado, M. L. (2016b). From the CLIL craze to the CLIL conundrum: Addressing the current CLIL controversy. *Bellaterra Journal of Teaching & Learning Language & Literature*, 9(1), 9. <https://doi.org/10.5565/rev/jtl3.667>
- Pérez Cañado, M. L. (2018). CLIL and pedagogical innovation: Fact or fiction? *International Journal of Applied Linguistics*, 28(3), 369–390. <https://doi.org/10.1111/ijal.12208>
- Pérez Cañado, M. L. (2018). The effects of CLIL on L1 and content learning: Updated empirical evidence from monolingual contexts. *Learning and Instruction*, 57, 18–33.  
<https://doi.org/10.1016/j.learninstruc.2017.12.002>
- Pérez Cañado, M. L. (2020). CLIL and elitism: Myth or reality? *The Language Learning Journal*, 48(1), 4–17. <https://doi.org/10.1080/09571736.2019.1645872>
- Pérez Cañado, M. L. (2021). Inclusion and diversity in bilingual education: A European comparative study. *International Journal of Bilingual Education and Bilingualism*(ahead of print), 1–17. <https://doi.org/10.1080/13670050.2021.2013770>
- Pérez Cañado, M. L., & Lancaster, N. K. (2017). The effects of CLIL on oral comprehension and production: A longitudinal case study. *Language, Culture and Curriculum*, 30(3), 300–316. <https://doi.org/10.1080/07908318.2017.1338717>
- Pérez Vidal, C. (2009). The integration of content and language in the classroom: A European approach to education (the second time around). In E. Dafouz (Ed.), *Richmond CLIL handbooks for teachers. CLIL across educational levels: Experiences from primary, secondary and tertiary contexts* (pp. 3–16). Richmond Publ [u.a.].
- Pérez-Vidal, C., & Roquet, H. (2015). The linguistic impact of a CLIL science programme: An analysis measuring relative gains. *System*, 54, 80–90.  
<https://doi.org/10.1016/j.system.2015.05.004>
- Piesche, N., Jonkmann, K., Fiege, C., & Keßler, J.-U. (2016). CLIL for all? A randomised controlled field experiment with sixth-grade students on the effects of content and language integrated science learning. *Learning and Instruction*, 44, 108–116.  
<https://doi.org/10.1016/j.learninstruc.2016.04.001>
- Pladevall-Ballester, E. (2015). Exploring primary school CLIL perceptions in Catalonia: Students', teachers' and parents' opinions and expectations. *International Journal of Bilingual Education and Bilingualism*, 18(1), 45–59.  
<https://doi.org/10.1080/13670050.2013.874972>
- Pladevall-Ballester, E. (2016). CLIL subject selection and young learners' listening and reading comprehension skills. *International Journal of Applied Linguistics*, 26(1), 52–74.  
<https://doi.org/10.1111/ijal.12079>
- Pladevall-Ballester, E. (2019). A longitudinal study of primary school EFL learning motivation in CLIL and non-CLIL settings. *Language Teaching Research*, 23(6), 765–786.  
<https://doi.org/10.1177/1362168818765877>

- Pladevall-Ballester, E., & Vallbona, A. (2016). CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills. *System*, 58, 37–48.  
<https://doi.org/10.1016/j.system.2016.02.009>
- Plomp, T. (2013). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research* (pp. 10–51). SLO.
- Plomp, T., & Nieveen, N. (Eds.). (2013). *Educational design research*. SLO.
- Priestley, M., & Biesta, G. (Eds.). (2017). *Reinventing the curriculum: New trends in curriculum policy and practice*. Bloomsbury Pub.
- Prieto-Arranz, J. I., Rallo Fabra, L., Calafat-Ripoll, C., & Catrain-González, M. (2015). Testing progress on receptive skills in CLIL and non-CLIL contexts. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Educational linguistics. Content-based language learning in multilingual educational environments* (Vol. 23, pp. 123–137). Springer.  
[https://doi.org/10.1007/978-3-319-11496-5\\_8](https://doi.org/10.1007/978-3-319-11496-5_8)
- Quintana Aguilera, J. A., Restrepo Castro, D., Romero, G., & Cárdenas Messa, G. A. (2019). The effect of content and language integrated learning on the development of English reading comprehension skills. *Lenguaje*, 47(2), 427–452.  
<https://doi.org/10.25100/lenguaje.v47i2.7699>
- Rallo Fabra, L., & Jacob, K. (2015). Does CLIL enhance oral skills? Fluency and pronunciation errors by Spanish-Catalan learners of English. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Educational linguistics. Content-based language learning in multilingual educational environments* (Vol. 23, pp. 163–177). Springer.  
[https://doi.org/10.1007/978-3-319-11496-5\\_10](https://doi.org/10.1007/978-3-319-11496-5_10)
- Rallo Fabra, L., & Juan-Garau, M. (2011). Assessing FL pronunciation in a semi-immersion setting: The effects of CLIL instruction on Spanish-Catalan learners' perceived comprehensibility and accentedness. *Poznań Studies in Contemporary Linguistics*, 47(1), 96–108. <https://doi.org/10.2478/psicl-2011-0008>
- Reinking, D., & Bradley, B. A. (Eds.). (2008). *Language and literacy series. On formative and design experiments: Approaches to language and literacy research*. Teachers College Press.
- Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction*, 30(1), 86–112.  
<https://doi.org/10.1080/07370008.2011.634081>
- Richter, K. (2019). *English-medium instruction and pronunciation*. Multilingual Matters.  
<https://doi.org/10.21832/RICHTE2456>
- Riemer, C. (2016). Befragung [inquiry]. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik [research methods in language teaching research]: Ein Handbuch [a handbook]* (1st ed., pp. 155–172). Narr Francke Attempto.
- Roberts, L., González Alonso, J., Pliatsikas, C., & Rothman, J. (2018). Evidence from neurolinguistic methodologies: Can it actually inform linguistic/language acquisition theories and translate to evidence-based applications? *Second Language Research*, 34(1), 125–143. <https://doi.org/10.1177/0267658316644010>

- Roiha, A. (2014). Teachers' views on differentiation in content and language integrated learning (CLIL): Perceptions, practices and challenges. *Language and Education*, 28(1), 1–18. <https://doi.org/10.1080/09500782.2012.748061>
- Roiha, A. (2019). Investigating former pupils' experiences and perceptions of CLIL in Finland: A retrospective analysis. *Nordic Journal of Studies in Educational Policy*, 5(2), 92–103. <https://doi.org/10.1080/20020317.2019.1586514>
- Roiha, A., & Mäntylä, K. (2021). CLIL as a vehicle for a positive English selfconcept: An analysis of one former student's life course. In K. R. Talbot, M.-T. Gruber, & R. Nishida (Eds.), *The psychological experience of integrating content and language* (pp. 55–72). Multilingual Matters. <https://doi.org/10.21832/9781788924306-009>
- Roquet, H., & Pérez-Vidal, C. (2017). Do productive skills improve in content and language integrated learning contexts? The case of writing. *Applied Linguistics*, 38(4), 489–511. <https://doi.org/10.1093/applin/amv050>
- Rose, D. (2009). Writing as linguistic mastery: The development of genre-based literacy pedagogy. In R. Beard, D. Myhill, & J. Riley (Eds.), *The SAGE handbook of writing development* (pp. 151–166). SAGE Publications. <https://doi.org/10.4135/9780857021069.n11>
- Rose, D. (2014). Genre in the Sydney school. In J. P. Gee & M. Handford (Eds.), *Routledge handbooks in applied linguistics. The Routledge handbook of discourse analysis* (1st ed., pp. 209–225). Routledge.
- Rose, D. (2018). Languages of schooling: Embedding literacy learning with genre-based pedagogy. *European Journal of Applied Linguistics*, 6(1), 59–89. <https://doi.org/10.1515/eujal-2017-0008>
- Rose, D., & Martin, J. R. (2012). *Learning to write, reading to learn: Genre, knowledge and pedagogy in the Sydney school. Equinox textbooks and surveys in linguistics*. Equinox.
- Rosi, F. (2018). Content-specific learning in CLIL. *Educazione Linguistica Language Education*, 7(1). <https://doi.org/10.30687/ELLE/2280-6792/2018/01/002>
- Rothery, J. (1994). *Write it right: Resources for literacy and learning. Write it right, resources for literacy and learning*. Disadvantaged Schools Program, Metropolitan East Region, NSW Dept. of School Education.
- Roussel, S., Joulia, D., Tricot, A., & Sweller, J. (2017). Learning subject content through a foreign language should not ignore human cognitive architecture: A cognitive load theory approach. *Learning and Instruction*, 52, 69–79. <https://doi.org/10.1016/j.learninstruc.2017.04.007>
- Ruiz de Zarobe, Y. (2007). CLIL in a bilingual community: similarities and differences with learning English as a foreign language. *VIEWS: Vienna English Working Papers*, 16(3), 47–52.
- Ruiz de Zarobe, Y. (2010). Written production and CLIL: An empirical study. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *AILA applied linguistics series. Language use and language learning in CLIL classrooms* (Vol. 7, pp. 191–210). John Benjamins. <https://benjamins.com/catalog/aals.7.10rui>

- Ruiz de Zarobe, Y., & Cenoz, J. (2015). Way forward in the twenty-first century in content-based instruction: Moving towards integration. *Language, Culture and Curriculum*, 28(1), 90–96. <https://doi.org/10.1080/07908318.2014.1000927>
- Ruiz de Zarobe, Y., & Zenotz, V. (2018). Learning strategies in CLIL classrooms: How does strategy instruction affect reading competence over time? *International Journal of Bilingual Education and Bilingualism*, 21(3), 319–331. <https://doi.org/10.1080/13670050.2017.1391745>
- Rumlich, D. (2016). *Evaluating bilingual education in Germany: CLIL students' general English proficiency, EFL self-concept and interest*. Peter Lang.
- Rumlich, D. (2020). Bilingual education in monolingual contexts: A comparative perspective. *The Language Learning Journal*, 48(1), 115–119. <https://doi.org/10.1080/09571736.2019.1696879>
- Rüsen, J. (1983). *Historische Vernunft [historical reason]: Grundzüge einer Historik [basics of historiography]*. Kleine Vandenhoeck-Reihe. Vandenhoeck & Ruprecht.
- Rüsen, J. (2004). Historical consciousness: Narrative structure, moral function, and ontogenetic development. In P. Seixas (Ed.), *Theorizing historical consciousness* (pp. 63–85). University of Toronto Press.
- Ryshina-Pankova, M. (2016). Scaffolding advanced literacy in the foreign language classroom: Implementing a genre-driven content-based approach. In L. Cammarata (Ed.), *Content-based foreign language teaching: Curriculum and pedagogy for developing advanced thinking and literacy skills* (pp. 51–76). Routledge.
- San Isidro, X. (2019). The multi-faceted effects of CLIL: A literature review. *Nexus*(1), 33–49.
- San Isidro, X., & Lasagabaster, D. (2019). The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research*, 23(5), 584–602. <https://doi.org/10.1177/1362168817754103>
- Sandoval, W. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences*, 23(1), 18–36. <https://doi.org/10.1080/10508406.2013.778204>
- Schall-Leckrone, L., & Barron, D. (2018). Apprenticing students and teachers into historical content, language, and thinking through genre pedagogy. In L. C. d. Oliveira & K. M. Obenchain (Eds.), *Teaching history and social studies to English language learners* (Vol. 16, pp. 205–231). Springer. [https://doi.org/10.1007/978-3-319-63736-5\\_9](https://doi.org/10.1007/978-3-319-63736-5_9)
- Schall-Leckrone, L., & McQuillan, P. J. (2012). Preparing history teachers to work with English learners through a focus on the academic language of historical analysis. *Journal of English for Academic Purposes*, 11(3), 246–266. <https://doi.org/10.1016/j.jeap.2012.05.001>
- Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, 12(4), 431–459. [https://doi.org/10.1016/S0898-5898\(01\)00073-0](https://doi.org/10.1016/S0898-5898(01)00073-0)
- Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Lawrence Erlbaum Associates.
- Schleppegrell, M. J., Achugar, M., & Oteiza, T. (2004). The Grammar of history: Enhancing content-based instruction through a functional focus on language. *TESOL Quarterly*, 38(1), 67–93.

- Schlepppegrell, M. J., Greer, S., & Taylor, S. (2008). Literacy in history: Language and meaning. *Australian Journal of Language and Literacy*, 31(2), Article 174-187.
- Schmidt, W. (Ed.). (1981). *Funktional-kommunikative Sprachbeschreibung [functional communicative description of language]*. Bibliographisches Institut.
- Schönemann, B. (2012). Geschichtsbewusstsein [historical consciousness]: Theorie [theory]. In M. Barricelli & M. Lücke (Eds.), *Handbuch Praxis des Geschichtsunterrichts [handbook of the practice of history teaching]* (pp. 98–111). Wochenschau Verlag.
- Schöner, A. (2007). Kompetenzbereich Historische Sachkompetenzen [factual competences]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 265–314). ars una.
- Schramm, K., & Schwab, G. (2016). Beobachtung [observation]. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik [research methods in language teaching research]: Ein Handbuch [a handbook]* (1st ed., pp. 141–154). Narr Francke Attempto.
- Schreiber, W. (n.d.). *FUER Geschichtsbewusstsein. Research Project* (A. Schöner, Trans.). <https://www1.ku.de/GGF/Didaktik/Projekt/English%20version/theory.html>
- Schreiber, W. (2007a). Kompetenzbereich Historische Fragekompetenzen [questioning competences]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 155–193). ars una.
- Schreiber, W. (2007b). Kompetenzbereich Historische Methodenkompetenzen [methodological competences]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 194–235). ars una.
- Schreiber, W. (2007c). Kompetenzbereich Historische Orientierungskompetenzen [orientation competences]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 236–264). ars una.
- Schreiber, W., Körber, A., Borries, B. v., Krammer, R., Leutner-Ramme, S., Mebus, S., Schöner, A., & Ziegler, B. (2007). Historisches Denken [historical thinking]: Ein Kompetenz-Strukturmodell (Basis Beitrag) [a structural model of competence (foundations)]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 17–53). ars una.
- Schreiber, W., Schöner, A., & Sochatzy, F. (2013). *Analyse von Schulbüchern als Grundlage empirischer Geschichtsdidaktik [analysis of course books as foundation of empirical history didactics]*. Kohlhammer.

- Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE Publications.
- Schreier, M. (2013). Qualitative content analysis. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 170–184). SAGE Publications.
- Schwarz, M. (2020). *Beyond the walls: A mixed methods study of teenagers' extramural English practices and their vocabulary knowledge* [PhD thesis]. University of Vienna, Vienna.
- Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139173438>
- Seixas, P. (2017). A model of historical thinking. *Educational Philosophy and Theory*, 49(6), 593–605. <https://doi.org/10.1080/00131857.2015.1101363>
- Seixas, P., & Morton, T. (2013). *The big six: Historical thinking concepts*. Nelson Education.
- Shavelson, R., Phillips, L., & Towne, M. (2003). On the science of education design studies. *Educational Researcher*, 32(1), 25.
- Shemilt, D. (1980). *Schools' council history 13-16 project*. Holmes-McDougall.
- Shemilt, D. (1983). The devil's locomotive. *History and Theory*, 22(4), 1–18.
- Siepmann, P., Rumlich, D., Matz, F., & Römhild, R. (2021). Attention to diversity in German CLIL classrooms: Multi-perspective research on students' and teachers' perceptions. *International Journal of Bilingual Education and Bilingualism*(ahead of print).
- Simons, M., Vanhees, C., Smits, T., & van de Putte, K. (2019). Remediating foreign language anxiety through CLIL? A mixed-methods study with pupils, teachers and parents. *Revista De Lingüística Y Lenguas Aplicadas*, 14(1), 153. <https://doi.org/10.4995/rlyla.2019.10527>
- Skinnari, K. (2020). CLIL challenges: Secondary school CLIL teachers' voices and experienced agency in three European contexts. *Journal for the Psychology of Language Learning*, 2(2), AOP.
- Skinnari, K., & Bovellan, E. (2016). CLIL teachers' beliefs about integration and about their professional roles: Perspectives from a European context. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Bilingual education and bilingualism. Conceptualising integration in CLIL and multilingual education* (pp. 145–170). Multilingual Matters.
- Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology*, 10, 2366. <https://doi.org/10.3389/fpsyg.2019.02366>
- Smit, U., & Finker, T. (2018). CLIL in Austrian technical colleges (HTL): An analysis of classroom practices based on systematic video-based lesson observations. In M. Dannerer & P. Mauser (Eds.), *Formen der Mehrsprachigkeit [forms of multilingualism]: Sprachen und Varietäten in sekundären und tertiären Bildungskontexten [languages and varieties in the context of secondary and tertiary education]* (pp. 229–246). Stauffenburg Verlag.
- Smit, U., & Schwarz, M. (2019). English in Austria: Policies and practices. In R. Hickey (Ed.), *English in the German-speaking world* (Vol. 25, pp. 294–314). Cambridge University Press. <https://doi.org/10.1017/9781108768924.015>
- Smith, F. (1992). *To think: In language, learning and education*. Teachers College Press.
- Snow, M. A., Met, M., & Genesee, F. (1989). A conceptual framework for the integration of language and content in second/foreign language instruction. *TESOL Quarterly*, 23(2), 201–217.

- Somers, T. (2017). Content and language integrated learning and the inclusion of immigrant minority language students: A research review. *International Review of Education*, 63(4), 495–520.
- Somers, T., & Llinares, A. G. (2018). Students' motivation for content and language integrated learning and the role of programme intensity. *International Journal of Bilingual Education and Bilingualism*, 1(4), 1–16.  
<https://doi.org/10.1080/13670050.2018.1517722>
- Staschen-Dielmann, S. (2010). Eine integrierte Beurteilung von fachspezifischen und fremdsprachlichen Kompetenzen [integrated assessment of subject-specific and linguistic competences]: Vorschläge für die Leistungsfeststellung im bilingualen Geschichtsunterricht [suggestions for assessment in bilingual history education]. In S. Dorff (Ed.), *Bilingualer Sachfachunterricht in der Sekundarstufe [bilingual content teaching in secondary education]: Eine Einführung [an introduction]* (pp. 228–241). Narr Francke Attempto.
- State Institute of Bavaria. (2005). *Operatorenliste Sozialkunde [list of performative verbs for social studies]*. [https://www.isb.bayern.de/download/21823/epa\\_operatorenliste\\_isb.pdf](https://www.isb.bayern.de/download/21823/epa_operatorenliste_isb.pdf)
- Statistik Austria. (2020). *Bildung in Zahlen 2018/2019 [education in numbers 2018/2019]*. Statistik Austria.
- Surmont, J., Struys, E., van den Noort, M., & van de Craen, P. (2016). The effects of CLIL on mathematical content learning: A longitudinal study. *Studies in Second Language Learning and Teaching*, 6(2), 319. <https://doi.org/10.14746/ssllt.2016.6.2.7>
- Surmont, J., van der Craen, P., Struys, E., & Somers, T. (2014). Evaluating a CLIL student: Where to find the CLIL advantage. In R. Breeze, C. Llamas Saíz, & C. Martínez Pasamar (Eds.), *Utrecht studies in language and communication. Integration of theory and practice in CLIL* (Vol. 28, pp. 55–74). Rodopi.
- Swain, m., Kinnear, P., & Steinman, L. (2015). *Sociocultural theory in second language education*. Multilingual Matters. <https://doi.org/10.21832/9781783093182>
- Swain, m., & Lapkin, S. (1989). Canadian immersion and adult second language teaching: What's the connection? *The Modern Language Journal*, 73(2), 150–159.  
<https://doi.org/10.1111/j.1540-4781.1989.tb02537.x>
- Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation. Cognition in education* (Vol. 55, pp. 37–76). Elsevier.  
<https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Sylvén, L. K., & Ohlander, S. (2019a). English reading comprehension. In L. K. Sylvén (Ed.), *Bilingual education & bilingualism. Investigating content and language integrated learning: Insights from Swedish high schools* (Vol. 116, pp. 136–151). Multilingual Matters.
- Sylvén, L. K., & Ohlander, S. (2019b). English receptive vocabulary. In L. K. Sylvén (Ed.), *Bilingual education & bilingualism. Investigating content and language integrated learning: Insights from Swedish high schools* (Vol. 116, pp. 101–116). Multilingual Matters.
- Talbot, K. R., Gruber, M.-T., Lämmerer, A., Hofstadler, N., & Mercer, S. (2021). Comparatively speaking: CLIL/EMI teacher well-being at the primary, secondary and tertiary levels in

- Austria. In K. R. Talbot, M.-T. Gruber, & R. Nishida (Eds.), *The psychological experience of integrating content and language* (pp. 153–173). Multilingual Matters.
- Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research*. Advance online publication. <https://doi.org/10.1177/1362168811401153>
- Teacher Training College Upper Austria. (2020). *Fortbildungsprogramm BMHS/ BS 2020/21 [in-service teacher training programme BMHS/ BS 2020/21]*. [https://ph-ooe.at/fileadmin/Daten\\_PHOOE/Fort-Weiterbildung\\_BS\\_BMHS/Fortbildung\\_2020\\_21/Fortbildungsprogramm\\_BMHS\\_2020.pdf](https://ph-ooe.at/fileadmin/Daten_PHOOE/Fort-Weiterbildung_BS_BMHS/Fortbildung_2020_21/Fortbildungsprogramm_BMHS_2020.pdf)
- Teacher Training College Vienna. (2020). *DLP-CLIL foundation of content and language integrated learning*. Teacher Training College Vienna. <https://www.phwien.ac.at/28-hochschullehrer-gaenge-und-fortbildungsangebot/lehrgaenge/496-710289dc-dlp-clil-foundation-of-content-and-language-integrated-learning>
- Tedick, D. J., & Lyster, R. (2019). *Scaffolding language development in immersion and dual language classrooms*. Routledge.
- Tedick, D. J., & Wesely, P. M. (2015). A review of research on content-based foreign/second language education in US K-12 contexts. *Language, Culture and Curriculum*, 28(1), 25–40. <https://doi.org/10.1080/07908318.2014.1000923>
- Tedick, D. J., & Young, A. I. (2018). Two-way immersion students' home languages, proficiency levels, and responses to form-focused instruction. *International Journal of Bilingual Education and Bilingualism*, 21(3), 303–318. <https://doi.org/10.1080/13670050.2017.1383354>
- Theis, R. (2010). Bilingualer Geschichtsunterricht [bilingual history education]. In S. Dorff (Ed.), *Bilingualer Sachfachunterricht in der Sekundarstufe [bilingual content teaching in secondary education]: Eine Einführung [an introduction]* (pp. 44–57). Narr Francke Attempto.
- Thompson, A. S., & Sylvén, L. K. (2015). "Does English make you nervous?" Anxiety profiles of CLIL and non-CLIL students in Sweden. *Apples - Journal of Applied Language Studies*, 9(2), 1–23. <https://doi.org/10.17011/apples/urn.201512093950>
- Thürmann, E. (2010). Zur Konstruktion von Sprachgerüsten im bilingualen Sachfachunterricht [the construction of language scaffolding in bilingual content teaching]. In S. Dorff (Ed.), *Bilingualer Sachfachunterricht in der Sekundarstufe [bilingual content teaching in secondary education]: Eine Einführung [an introduction]* (pp. 137–153). Narr Francke Attempto.
- Tomlinson, C. A. (2001). *How to differentiate instruction in mixed ability classrooms* (2nd ed.). ASCD. <http://gbv.ebib.com/patron/FullRecord.aspx?p=280341>
- Trautwein, U., Bertram, C., Borries, B. v., Brauch, N., Hirsch, M., Schröter, K., Körber, A., Kühberger, C., Meyer-Hamme, J., Merkt, M., Neureiter, H., Schwan, S., Schreiber, W., Wagner, W., Waldis, M., Werner, M., Ziegler, B., & Zuckowski, A. (2017). *Kompetenzen historischen Denkens erfassen [assessing competences of historical thinking]: Konzeption, Operationalisierung und Befunde des Projekts „Historical Thinking – Competencies in History“ (HiTCH) [conception, operationalisation, and evidence of the project "historical*



- thinking - competencies in history" (HiTCH)]*. Waxmann Verlag. [http://www.content-select.com/index.php?id=bib\\_view&ean=9783830985983](http://www.content-select.com/index.php?id=bib_view&ean=9783830985983)
- Trimble, L. (1985). *English for science and technology: A discourse approach*. Cambridge language teaching library. Cambridge University Press.
- Ur, P. (2012). *A course in English language teaching* (2nd rev. ed.). Cambridge University Press.
- van Borries, B. (2007). Empirie [empirical evidence]: Ergebnisse messen [measuring outcomes]. In A. Körber, W. Schreiber, & A. Schöner (Eds.), *Kompetenzen historischen Denkens [competences of historical thinking]: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik [a structural model contributing to competency-based history didactics]* (pp. 653–673). ars una.
- van Boxtel, C., & van Drie, J. (2004). Historical reasoning: A comparison of how experts and novices contextualise historical sources. *History Education Research Journal*, 4(2), 89–97. <https://doi.org/10.18546/herj.04.2.10>
- van den Akker, J. (1999). Principles and methods of development research. In J. van den Akker, N. Nieveen, T. Plomp, K. Gustafson, & R. M. Branch (Eds.), *Design approaches and tools in education and training* (pp. 1–14). Springer Netherlands. <http://nbn-resolving.de/urn:nbn:de:1111-201108211273>
- van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006a). *Educational design research*. Taylor & Francis Ltd.
- van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006b). Introducing educational design research. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 3–7). Taylor & Francis Ltd.
- van den Akker, J., & Nieveen, N. (2016). The role of teachers in design research in education. In S. Doff & R. Komoss (Eds.), *Making change happen: Wandel im Fachunterricht analysieren und gestalten [analysing and designing change in content teaching]*. Springer.
- van den Akker, J., Nieveen, N., Plomp, T., Gustafson, K., & Branch, R. M. (Eds.). (1999). *Design approaches and tools in education and training*. Springer Netherlands.
- van der Walt, C., & Ruiters, J. (2012). Every teacher a language teacher? Developing awareness of multilingualism in teacher education. *Journal for Language Teaching*, 45(2). <https://doi.org/10.4314/jlt.v45i2.5>
- van Drie, J., & van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review*, 20(2), 87–110. <https://doi.org/10.1007/s10648-007-9056-1>
- van Kampen, E., Admiraal, W., & Berry, A. (2018). Content and language integrated learning in the Netherlands: Teachers' self-reported pedagogical practices. *International Journal of Bilingual Education and Bilingualism*, 21(2), 222–236. <https://doi.org/10.1080/13670050.2016.1154004>
- van Kampen, E., Meirink, J., Admiraal, W., & Berry, A. (2017). Do we all share the same goals for content and language integrated learning (CLIL)? Specialist and practitioner perceptions of 'ideal' CLIL pedagogies in the Netherlands. *International Journal of Bilingual Education and Bilingualism*, 4, 1–17. <https://doi.org/10.1080/13670050.2017.1411332>
- van Mensel, L., Hilgsmann, P., Mettwie, L., & Galand, B. (2020). CLIL, an elitist language learning approach? A background analysis of English and Dutch CLIL pupils in French-speaking

- Belgium. *Language, Culture and Curriculum*, 33(1), 1–14.  
<https://doi.org/10.1080/07908318.2019.1571078>
- Vanderbeke, M., & Wilden, E. (2017). Sachfachliche Diskursfähigkeit durch fremdsprachliche affordances in bilingualen Schülerlaborprojekten [subject-specific discourse skills through linguistic affordances in bilingual learner projects]. *Zeitschrift Für Fremdsprachenforschung*, 28(1), 3–27.
- Verband der Geschichtslehrer Deutschlands. (2006). *Bildungsstandards Geschichte [educational standards history]: Rahmenmodell Gymnasium 5. - 10. Jahrgangsstufe [framework for academic schools, grade 5-10]. Studien des Verbands der Geschichtslehrer Deutschlands*. Wochenschau Verlag.
- VERBI Software. (2017). *MaxQDA 2018* [Computer software]. Berlin. maxqda.com
- VERBI Software. (2019). *MaxQDA 2020* [Computer software]. Berlin. maxqda.com
- Verspoor, M., Bot, K. de, & Xu, X. (2015). The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education*, 3(1), 4–27.  
<https://doi.org/10.1075/jicb.3.1.01ver>
- Vienna Board of Education. (2016). *Modern language initiatives*.  
[http://www.schulentwicklung.at/joomla/images/stories/Sprachinitiativen/Fremdsprachenmodell\\_SJ\\_16\\_17.pdf](http://www.schulentwicklung.at/joomla/images/stories/Sprachinitiativen/Fremdsprachenmodell_SJ_16_17.pdf)
- Vienna Board of Education. (2020). *Vienna bilingual schooling (VBS)*. <https://www.bildung-wien.gv.at/schulen/Oesterreichisches-Schulsystem/Berufsbildende-mittlere-und-here-Schulen--BMHS-/Vienna-Bilingual-Schooling.html>
- Villarreal Olaizola, I., & García Mayo, M. d. P. (2009). Tense and agreement morphology in the interlanguage of Basque/Spanish bilinguals: CLIL versus non-CLIL. In Y. d. Ruiz Zarobe & R. M. Jiménez Catalán (Eds.), *Second language acquisition. Content and language integrated learning: Evidence from research in Europe* (Vol. 41, pp. 157–175). Multilingual Matters.
- Vollmer, H. J. (2010). Fachkompetenz als fachbasierte Diskursfähigkeit am Beispiel Geographie [subject competence as subject-specific discourse competence in the case of geography teaching]. In S. Dorff (Ed.), *Bilingualer Sachfachunterricht in der Sekundarstufe [bilingual content teaching in secondary education]: Eine Einführung [an introduction]* (pp. 242–257). Narr Francke Attempto.
- Vollmer, H. J. (2011). *Schulsprachliche Kompetenzen [competences of the language of schooling]: Zentrale Diskursfunktionen [central discourse functions]*.  
<file:///C:/Users/43699/AppData/Local/Temp/VollmerDF-Kurzdefinitionen.pdf>
- Vollmer, H. J., & Thürmann, E. (2010). Zur Sprachlichkeit des Fachlernens [the linguistic dimension of content learning]: Modellierung eines Referenzrahmens für Deutsch als Zweitsprache [modelling a reference framework for German as a second language]. In B. Ahrenholz (Ed.), *Fachunterricht und Deutsch als Zweitsprache [subject teaching and German as second language]* (2nd ed., pp. 107–132). Narr Francke Attempto.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (A. Blunden & N. Schmolze, Trans.) (rev. ed.). Harvard University Press.
- Vygotsky, L. S. (1987). *Thought and language* (A. Kozulin, Trans.) (rev. ed.). MIT Press.

- Walker, E. (2010). A systemic functional contribution to planning academic genre teaching in a bilingual education context. *Language Awareness*, 19(2), 73–87.  
<https://doi.org/10.1080/09658410903431721>
- Walsh, B. (2002). *Modern world history* (2nd repr. ed.). *History in focus*. Murray.
- Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53(4), 5–23.  
<https://doi.org/10.1007/BF02504682>
- Weinert, F. E. (2001). Competencies and key competencies: Educational perspective. In Smelser, Neil; Baltes, Paul (Ed.), *International encyclopedia of the social and behavioral* (pp. 2433–2436). Elsevier Science.
- Weinert, F. E. (Ed.). (2002). *Leistungsmessungen in Schulen [measuring performance in schools]* (2nd ed.). Beltz.
- Wesely, P. M. (2012). Learner attitudes, perceptions, and beliefs in language learning. *Foreign Language Annals*, 45(1), 98–114.
- Whittaker, R. (2018). Reading to learn in CLIL subjects: Working with content-language. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, 1(1), 19.  
<https://doi.org/10.5565/rev/clil.4>
- Whittaker, R., & García Parejo, I. (2018). Teacher learning for European literacy education (TeL4ELE): Genre-based pedagogy in five European countries. *European Journal of Applied Linguistics*, 6(1), 31–57. <https://doi.org/10.1515/eujal-2017-0021>
- Whittaker, R., Llinares, A., & McCabe, A. (2011). Written discourse development in CLIL at secondary school. *Language Teaching Research*, 15(3), 343–362.  
<https://doi.org/10.1177/1362168811401154>
- Whittaker, R., & McCabe, A. (2020). Expressing evaluation across disciplines in primary and secondary CLIL writing: A longitudinal study. *International Journal of Bilingual Education and Bilingualism*, 5(2), 1–18. <https://doi.org/10.1080/13670050.2020.1798869>
- Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford University Press.
- Wineburg, S. S. (1991). On the reading of historical texts: Notes on the breach between school and academy. *American Educational Research Journal*, 28(3), 495–519.  
<https://doi.org/10.3102/00028312028003495>
- Xanthou, M. (2011). The impact of CLIL on L2 vocabulary development and content knowledge. *English Teaching: Practice and Critique*, 10(4), 116–126.
- Zhyrun, I. (2016). Culture through comparison: Creating audio-visual listening materials for a CLIL course. *Latin American Journal of Content and Language Integrated Learning*, 9(2), 345–373. <https://doi.org/10.5294/laclil.2016.9.2.5>
- Zydatiś, W. (2007). *Deutsch-Englische Züge in Berlin (DEZIBEL) [German-English streams in Berlin]: Eine Evaluation des bilingualen Sachfachunterrichts an Gymnasien [an evaluation of bilingual subject teaching at academic secondary schools]*. Peter Lang.

## Deutsche Zusammenfassung

Content and Language Integrated Learning (CLIL) versteht sich als ein Lehr-Lernansatz, bei dem Sprach- und Fachlernen integriert werden mit dem ursprünglichen Ziel, Sprachenlernen effizienter zu gestalten. Dementsprechend ist es wenig überraschend, dass CLIL vor allem aus der Perspektive der angewandten Linguistik beforscht wurde und dass das fachliche Lernen erst in letzter Zeit verstärkt beforscht wurde. Damit einhergehend entwickelte sich die Frage nach der Integration von Sprach- und Fachlernen zu einem zentralen Thema in der CLIL Forschung. Auf konzeptueller Ebene gibt es hier einige Vorschläge, welche sich meist im Bereich der Systemischen Funktionslinguistik und/ oder der soziokulturellen Theorie verorten lassen. Diese Ansätze führen zu sehr interessanten Einsichten bezüglich der Zusammenhänge zwischen Fach- und Sprachlernen, jedoch lassen sie sich nur schwer in der Unterrichtspraxis umsetzen. Ein Konzept, welches sowohl eine konzeptuelle Integration von Sprach- und Fachlernen zulässt, aber auch aus Sicht der Praxis greifbar und nützlich erscheint, ist das Konstrukt kognitiver Diskursfunktionen (*cognitive discourse functions, CDFs*; Dalton-Puffer, 2013). CDFs sind sprachliche Muster, die man routinemäßig zur Verbalisierung kognitiver Prozesse mit bestimmten Kommunikationsabsichten verwendet und stellen als solche einen essenziellen Bestandteil von Lehr- und Lernprozessen dar. Auch im Bereich der Geschichtsdidaktik konnte bereits gezeigt werden, dass CDFs eng mit zentralen fachlichen Kompetenzen verbunden sind, sowohl aus konzeptueller als auch empirischer Perspektive. Allerdings wurde dieses Konstrukt bisher noch nicht für den didaktischen Einsatz aufbereitet bzw. bräuchte es grundsätzlich mehr Forschung, die sich mit den Verknüpfungen von sprach-und-fach-integrativem Lernen, pädagogischer Praxis und didaktischem Material auseinandersetzt. Dies wäre auch deshalb wichtig, da es Lehrpersonen an sprach-und-fach-integrativem Material sowie dessen konzeptuellem Verständnis oftmals mangelt.

Um dieser Forschungslücke entgegenzuwirken, wurde in der vorliegenden Studie Design-Based Research (DBR) als Forschungsansatz gewählt. Dieser transdisziplinären Forschungsmethodik wird in der Literatur oftmals das Potential zugeschrieben, durch einen dualen Ansatz Theorie und Praxis erfolgreich verbinden zu können. Somit setzt sich diese Studie das Ziel, das theoretische Fundament der Sprach- und Fachintegration näher zu beleuchten, aber auch praxis-orientierte Ressourcen und Materialien für den CLIL Geschichtsunterricht der Sekundarstufe II zu erstellen, welche Fach- und Sprachlernen verbinden. Um diesen Zielen gerecht zu werden, wurden gemeinsam mit Lehrkräften der Sekundarstufe CDF-basierte und kompetenzorientierte Geschichtsmaterialien entwickelt. Dazu wurden zuerst die Bedürfnisse der Teilnehmenden durch Einzelinterviews mit der jeweiligen Lehrkraft, Gruppeninterviews mit Schüler\*innen sowie durch schriftliche Aufgabenstellungen erhoben. Die daraus gewonnenen Einsichten wurden im Designprozess und bei der Materialentwicklung berücksichtigt. Das vorläufige Material wurde von der jeweiligen Lehrperson im regulären CLIL-Unterricht eingesetzt und aus der Sicht der

Lehrperson und der Schüler\*innen sowie durch wiederholte schriftliche Aufgabenstellungen evaluiert. Basierend auf diesen Ergebnissen wurden der entwickelte Lehransatz und die erstellten Materialien in drei solcher Forschungszyklen in zwei verschiedenen Kontexten weiterentwickelt.

Die Ergebnisse dieser Studie deuten darauf hin, dass CDFs einen ökologisch-validen und effektiven Zugang bieten, um Fach- und Sprachlernen zu verbinden. Damit CDF-basierte Materialien von der Zielgruppe angenommen werden und zu positiven Lernerfolgen führen können, müssen aber eine Reihe von Bedingungen erfüllt werden. Ein wichtiger Aspekt ist beispielsweise, dass die Aufgabenstellungen einen (inter-)aktiven, abwechslungsreichen Lernprozess ermöglichen, der wiederum kleinschrittig aufbereitet wurde (*Scaffolding*). Darüber hinaus sollten die sprachlichen Unterstützungsmaßnahmen und deren Bestandteile immer aus der Sicht der Fachdisziplin betrachtet werden. Folglich erscheint es ebenso wichtig, nicht nur sprachliche Formen und Funktionen zu berücksichtigen, sondern auch fachspezifische und fachlich relevante Konzepte. Außerdem hat sich im Laufe des Projekts gezeigt, dass Methoden der Binnendifferenzierung eine zentrale Rolle im subjektiven Erfahren der Lernenden spielen.

Diese Aspekte stellten sich als ausschlaggebend für die Akzeptanz des neuen Lehransatzes bei den Teilnehmenden heraus, was sich wiederum in den Performanzen der Lernenden widerspiegelte. Vor den Interventionen schien es beiden Kohorten der Hauptstudie schwer zu fallen, fachliche Kompetenzen adäquat in der Fremdsprache umzusetzen. Beispielsweise hatten viele CLIL Schüler\*innen Probleme, Begründungen anzuführen, Kommunikationsabsichten zu signalisieren oder ihre Ideen angemessen zu verbinden. Im Falle von Gruppe A, welche zwei Interventionen durchlief, verbesserte sich im Laufe der Studie die Beurteilung der schriftlichen Performanzen sowohl in Hinblick auf akademische Sprache als auch fachspezifische Kompetenzen signifikant. Interessanterweise konnte im zweiten Durchgang ein erheblich größerer Leistungssprung beobachtet werden. Die Ergebnisse der Gruppe B, welche nur an einem Forschungszyklus teilnahm, verbesserten sich dagegen nur im sprachlichen Bereich, während die fachliche Dimension unverändert blieb. Abschließend konnte durch die Studie gezeigt werden, dass das CDF Konstrukt ein praktisches und überschaubares Forschungswerkzeug darstellt. Um allerdings eine verlässliche Codierung sicherstellen zu können, wären weitere Spezifizierungen aus Sicht der einzelnen Disziplinen notwendig. Die vorliegende Arbeit bietet diesbezüglichen einen Vorschlag für das Fach Geschichte.

## Digital appendix

All appendices are permanently stored in *Phaidra*, an online repository providing permanent and safe storage of digital files, run by the University of Vienna.

This is the link to the complete collection connected to this thesis:

<https://phaidra.univie.ac.at/o:1411771>

In this main collection, you will find a number of sub-collections. Here are the links for the different sub-collections:

- I. Instruments: <https://phaidra.univie.ac.at/o:1411629>
  - A. Interview guides: <https://phaidra.univie.ac.at/o:1411610>
  - B. Prompts for pre- & post-tasks: <https://phaidra.univie.ac.at/o:1411615>
  - C. Codebooks and code trees: <https://phaidra.univie.ac.at/o:1411624>
  - D. Rubrics: <https://phaidra.univie.ac.at/o:1422961>
  - E. Feedback sheet for students: <https://phaidra.univie.ac.at/o:1411626>
  - F. Transcription rules: <https://phaidra.univie.ac.at/o:1411628>
- II. Data analysis: <https://phaidra.univie.ac.at/o:1411698>
  - A. Interviews: <https://phaidra.univie.ac.at/o:1411653>
    - 1. Needs analysis: <https://phaidra.univie.ac.at/o:1411650>
    - 2. Evaluation unit I: <https://phaidra.univie.ac.at/o:1411651>
    - 3. Evaluation unit II: <https://phaidra.univie.ac.at/o:1411652>
  - A. Pre- and post-intervention tasks: <https://phaidra.univie.ac.at/o:1411689>
    - 1. Pilot cycle: <https://phaidra.univie.ac.at/o:1411656>
    - 2. Cycle 1: <https://phaidra.univie.ac.at/o:1411671>
    - 3. Cycle 2: <https://phaidra.univie.ac.at/o:1411677>
    - 4. Cycle 3: <https://phaidra.univie.ac.at/o:1411682>
    - 5. Needs analysis combined: <https://phaidra.univie.ac.at/o:1411687>
    - 6. Intrarater analysis: <https://phaidra.univie.ac.at/o:1411688>
  - B. Design sessions: <https://phaidra.univie.ac.at/o:1411694>
  - C. Meta data and overview: <https://phaidra.univie.ac.at/o:1411697>
- III. Didactic materials: <https://phaidra.univie.ac.at/o:1411764>
  - A. Unit I: absolutism & mercantilism: <https://phaidra.univie.ac.at/o:1411731>
  - B. Unit II: the Industrial Revolution: <https://phaidra.univie.ac.at/o:1411747>
  - C. Pilot units: ideologies of the 19<sup>th</sup> century: <https://phaidra.univie.ac.at/o:1411763>
- IV. Informed consent: <https://phaidra.univie.ac.at/o:1411770>

In hard-copy editions, please find the USB-drive at the back of the cover. This USB-drive follows the same structure as the online appendix.