universität wien

DISSERTATION

Titel der Dissertation

„Linguistische Analyse und korpusbasierte Extraktion deutscher und russischer wissenshaltiger Kontexte - Perspektiven für die Terminologiepraxis und Terminologiewissenschaft“

Verfasserin

Diplom-Übersetzerin Anne-Kathrin Schumann

angestrebter akademischer Grad

Doktorin der Philosophie (Dr. Phil.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 792 323
Dissertationsgebiet lt. Studienblatt:	Transkulturelle Kommunikation
Betreut von:	Prof. Dr. Gerhard Budin

„[... M]eaning is in essence *conceptual* ..., but is most easily studied through language.”
Cruse (2000: 32)

„At first glance the problem of definitions in natural language seems to be a more or less marginal question. This is, however, far from being true. While discussing definitions we shall be forced to touch upon some fairly intricate and central problems of linguistic theory. Though even the formal properties of definitional sentences are far from being clear, the more difficult problems arise with respect to their semantic interpretation, their role in introducing new terms into a given language, and their relation to non-definitional generic sentences.”
Bierwisch/Kiefer (1969: 55)

Inhalt

1. Einleitung 1
1.1 Zielstellung und Einordnung 1
1.2 Terminologiewissenschaftliche Perspektiven 3
1.3 Bedeutung wissenshaltiger Kontexte für die terminologische Praxis 5
1.4 Vorgehensweise und Gliederung 9

2. Wissenshaltige Kontexte 11
2.1 Terminologisch-methodologische Vorbemerkung 11
2.2 Kontexttypen in der Linguistik und erste Beschreibung sowie Abgrenzung wissenshaltiger Kontexte von anderen Kontexttypen 12
2.3 Wissenshaltige Kontexte: Begriffsdefinition und Klassifikation 14
2.3.1 Inhaltliche Definition 14
2.3.2 Klassifikation wissenshaltiger Kontexte 16
2.3.2.1 Formale Klassifikation 16
2.3.2.2 Epistemologische Klassifikation 17
2.3.2.3 Linguistische Klassifikation 19
2.3.3 Durch wissenshaltige Kontexte verbalisierte semantische Relationen 22
2.3.3.1 Eigenschaften semantischer Relationen und Bedeutung für die Terminologiearbeit 22
2.3.3.2 Entwicklung einer Typologie semantischer Relationen für die computergestützte Terminologiearbeit und die Identifikation wissenshaltiger Kontexte 24
2.4 Zusammenfassung 28

3. Wissenshaltige Kontexte: Stand der Forschung 29
3.1 Allgemeiner Forschungsstand: disziplinäre Eingrenzung 29
3.2 Linguistische Merkmale wissenshaltiger Kontexte 31
3.2.1 Definitionen und wissenshaltige Kontexte als sublanguage 31
3.2.2 Weitere linguistische Merkmale 32
3.2.3 Knowledge patterns 35
3.2.4 Zusammenfassung 43
3.3 Annotation wissenshaltiger Kontexte in Textkorpora 44
3.3.1 Durchführbarkeit der Annotation und Einflussfaktoren 44
3.3.2 Annotationsmethodologien 46
3.3.2.1 Annotationsverfahren 46
3.3.2.2 Evaluation manueller Annotationen 50
3.3.3 Zusammenfassung 52
3.4 Extraktion wissenshaltiger Kontexte 53
3.4.1 Symbolische Verfahren 53
3.4.2 Maschinelles Lernen 57
3.4.2.1 Klassifikationsverfahren 57
3.4.2.2 Regelerwerb 60

3.4.3 Hybride Ansätze *64*
3.4.4 Evaluation von Extraktionsergebnissen und Performanz von Extraktionsverfahren *69*
3.5 Zusammenfassung 70

4. Vorbereitung: Korpuserstellung und Annotation 72
4.1 Korpuserstellung 72
4.1.1 Babouk *72*
4.1.2 Lexikographische Korpora und Web-Korpora *72*
4.1.3 Dokumentation der erstellten Korpora *76*
4.1.4 Statistische Auswertung der erstellten Kopora *78*
4.1.4.1 Vergleich der Entwicklungs- und Testkorpora mit Referenzkorpora 78
4.1.4.2 Vergleich des Entwicklungskorpus‘ mit dem Testkorpus 84
4.2 Annotationsstudie zur Erstellung der Datenbasis 86
4.2.1 Ausgangsdaten *86*
4.2.2 Annotationsszenario und Annotation *88*
4.2.3 Ergebnisse und Bewertung *90*
4.2.3.1 Deutsch 90
4.2.3.2 Russisch 91
4.2.3.3 Quantitative Bewertung der Annotation 92
4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards 94
4.2.3.4.1 Identifikation der falschen Negative 95
4.2.3.4.2 Identifikation der falschen Positive 99
4.2.3.4.3 Goldstandard 100
4.2.4 Abdeckung der Annotation *100*
4.2.5 Zusammenfassung *101*
4.3 Analyse des Goldstandards 102
4.3.1 Theoriegeleitete Klassifikation von Textsorten in Web-Korpora *102*
4.3.2 Eigenschaften der Subkorpora *106*
4.4 Vorverarbeitung, Annotation und Formatierung der Korpora 109
4.4.1 Vorverarbeitung *109*
4.4.2 Linguistische Annotation *112*
4.4.2.1 Wortartentagging und Lemmatisierung 112
4.4.2.2 Syntaktische Analyse 113
4.4.2.2.1 Deutsch 114
4.4.2.2.2 Russisch 116
4.4.3 Formatierung *118*
4.5 Zusammenfassung 121

5. Linguistische Eigenschaften wissenshaltiger Kontexte 123
5.1 Einführung 123
5.2 Linguistische Fragestellungen 124
5.2.1 Ein Merkmalskatalog *124*
5.2.1.1 Aus der relevanten Literatur entnommene Merkmale und Hypothesen 124
5.2.1.2 Generalisierung 126
5.2.1.3 Systemisch-funktionale Linguistik 127

5.2.2 Lexikalische Merkmale 130
5.2.2.1 Methode 130
5.2.2.2 Wortartenfrequenzen 132
5.2.2.2.1 Russisch 132
5.2.2.2.2 Deutsch 135
5.2.2.3 Boost- und Stopwörter 139
5.2.2.3.1 Russisch 139
5.2.2.3.2 Deutsch 150
5.2.2.4 Zwischenfazit 156
5.2.2.5 Bausteine für „knowledge patterns“ 158
5.2.2.5.1 Russisch 158
5.2.2.5.2 Deutsch 162
5.2.2.5.3 Semantische Typen von „knowledge patterns“ 167
a) Prozesstypen und Verbklassen 167
b) Methode 170
c) Ergebnisse 171
5.2.2.6 Zwischenfazit 175
5.2.3 Strukturelle und grammatische Merkmale 176
5.2.3.1 Methode 176
5.2.3.2 Verbformenanalyse 179
5.2.3.2.1 Russisch 179
5.2.3.2.2 Deutsch 184
5.2.3.2.3 Zwischenfazit 186
5.2.3.3 Analyse des Definiendums 187
5.2.3.3.1 Russisch 187
5.2.3.3.2 Deutsch 190
5.2.3.3.3 Zwischenfazit 193
5.2.3.4 Syntaktische Komplexität 194
5.2.3.5 Satzlänge 195
5.2.3.5.1 Russisch 195
5.2.3.5.2 Deutsch 196
5.2.3.5.3 Zwischenfazit 196
5.2.3.6 Phraseologie 197
5.2.3.6.1 Russisch 197
5.2.3.6.2 Deutsch 199
5.2.3.6.3 Zwischenfazit 200
5.2.4 Zusammenfassung 200
5.3 Multivariate Analyse deutscher KRCs 202
5.3.1 Einführung 202
5.3.2 Logistische Regression 203
5.3.3 Merkmale und Daten 203
5.3.4 Modellierung und Ergebnisse 206
5.3.4.1 Variablenauswahl 207
5.3.4.2 Evaluation der Modellanpassung 208
5.3.4.3 Interpretation der Regressionskoeffizienten 208
5.3.4.4 Einfaches Klassifikationsmodell 211
5.3.4.5 Zusammenfassung 213
5.4 Fazit 214

6. Extraktion wissenshaltiger Kontexte 216
6.1 Einführung 216
6.2 Musterbasierte Extraktion wissenshaltiger Kontexte 217
6.2.1 Implementierung der Extraktionsmuster 217
6.2.2 Genutzte Muster 220
6.2.2.1 Deutsch 220
6.2.2.2 Russisch 221
6.2.3 Ergebnisse 222
6.2.3.1 Deutsch 222
6.2.3.1.1 Baseline 222
6.2.3.1.2 Punktuelle Präzisionsverbesserung 226
6.2.3.1.3 Möglichkeiten der punktuellen Abdeckungsverbesserung 227
6.2.3.2 Russisch 231
6.2.3.2.1 Baseline 231
6.2.3.2.2 Termerkennung 234
6.2.3.2.3 Zusätzliche Extraktionsmuster 236
6.2.3.2.4 Punktuelle Präzisionsverbesserung 236
6.2.4 Diskussion 239
6.3 Maschinelles Lernen 242
6.3.1 Russisch 243
6.3.1.1 Verringerung des Suchraums durch Vorsortierung der Korpussätze 243
6.3.1.1.1 Grundsätzliche Vorgehensweise 243
6.3.1.1.2 Daten und Modellierung 244
6.3.1.1.3 Ergebnisse 247
6.3.1.2 Einbezug der Musterinformation in die logistische Modellierung 250
6.3.1.2.1 Grundsätzliche Vorgehensweise 250
6.3.1.2.2 Daten und Modellierung 250
6.3.1.2.3 Ergebnisse 253
6.3.1.3 Zwischenfazit 256
6.3.2 Deutsch 257
6.3.2.1 Daten und Modellierung 257
6.3.2.2 Ergebnisse 260
6.3.2.3 Zwischenfazit 263
6.4 Qualität automatisch extrahierter wissenshaltiger Kontexte aus Web-Korpora und Perspektiven der Anreicherung von Terminologiedatenbanken mit korpusbasierter Information 263
6.5 Fazit 268

7. Zusammenfassung und Ausblick 270
7.1 Ergebnisse der Arbeit 270
7.2 Quo vadis, korpusbasierte Terminologie? 274
7.2.1 Terminologiepraktische Aspekte 275
7.2.2 Terminologiewissenschaftliche Aspekte 276

Bibliographie

Anhang

Tabellenverzeichnis

Abbildungsverzeichnis

Abstract

Lebenslauf der Autorin

1. Einleitung

1.1 Zielstellung und Einordnung

Das Ziel dieser Arbeit ist die Entwicklung einer Methode zur computergestützten Extraktion semantischen Wissens in Form wissenshaltiger Kontexte aus Fachtextkorpora. Im Einzelnen beschäftigt sich vorliegende Studie dabei mit den folgenden Fragestellungen:

- Wie wird fachspezifisches Wissen in Texten dargestellt, d. h. welche linguistischen Merkmale weisen wissenshaltige Kontexte auf?
- Welche Arten von fachspezifischem Wissen sind relevant für die übersetzungsorientierte Terminologiearbeit, d. h. welche Arten fachspezifischen Wissens sind in Terminologiedatenbanken darzustellen, um den Nutzern solcher Ressourcen einen effizienten und strukturierten, übersetzungsorientierten Wissenserwerb zu ermöglichen?
- Wie erkennen Übersetzer fachspezifisches Wissen in Texten, d. h. gibt es überhaupt einen Konsens in Bezug auf die Relevanz fachspezifischer Aussagen in Texten für die übersetzungsorientierte Terminologiearbeit?
- Wie kann relevantes semantisches Wissen in natürlichsprachigen Texten mit Hilfe von Computerprogrammen erkannt und automatisch aus ihnen extrahiert werden?
- Inwieweit sind automatisch aus Korpora extrahierte wissenshaltige Kontexte für die Terminologie überhaupt nutzbar?

Damit bearbeitet diese Dissertation ein Problemfeld, das am Schnittpunkt computerlinguistischer, korpuslinguistischer und terminologischer Teildisziplinen liegt. Insbesondere computerlinguistische Teilgebiete (in erster Linie Arbeiten zur Extraktion definitorischer Aussagen sowie von semantischen Relationen aus natürlichsprachigen Texten innerhalb der mitunter wohl nicht absolut trennscharf voneinander abgegrenzten Disziplinen *ontology learning* und *text mining*) sind dabei als Bezugspunkte zu nennen und werden im Verlauf dieser Arbeit in Form von Literaturverweisen an mehreren Stellen auftauchen. Zwar verfolgen Forschungsarbeiten in diesen Bereichen ein anderes Ziel als die nach wie vor auf den menschlichen Übersetzer ausgerichtete Terminologiearbeit, nämlich im weitesten Sinne die Entwicklung intelligenter, wissensbasierter Systeme (vgl. Ahmad 2001: 811 f.) bzw. allgemein das zielgerichtete Durchsuchen großer Textmengen. Inhaltliche und z. T. auch methodologische Überschneidungen treten aus terminologischer Sicht indes gerade im Hinblick auf die oben genannten Fragestellungen klar zutage. Als erster zentraler Begriff für das hier interessierende Forschungsgebiet lässt sich dabei der *Erwerb semantischen Wissens* ausmachen, womit das Bestreben bezeichnet wird, anhand der Analyse sprachlicher Artefakte Wissensressourcen für intelligente – menschliche oder maschinelle – Nutzer zu schaffen, so dass dieser Forschungszweig zusammen mit dem Erwerb *sprachlichen* Wissens aus Texten das allgemeinere Feld der *knowledge acquisition* bildet (ebd.: 810, Hervorhebung im Original):

> „The term *knowledge acquisition* assumes different shades of meaning as we cross the boundaries of various related disciplines. Computational linguists, corpus linguists and computational lexicographers use the term to refer to methods and techniques that enable humans to acquire knowledge of a given language from archives of language use."

Ahmads Zitat benennt auch den zweiten zentralen Begriff, der Forschungsarbeiten in dem hier interessierenden Fachgebiet zugrunde liegt: *archives of language use* oder Textkorpora, aus denen semantisches Wissen extrahiert wird. Betrachtet wird der Text – das Korpus – damit freilich weniger als linguistisch-sprachliche Einheit, denn als Kommunikationsmedium und Informationsträger, so dass der Terminus „Erwerb semantischen Wissens“ auf die kommunikative Funktion von Texten, die als Medien zur Weitergabe von relevantem Wissen genutzt werden, verweist. Studien in diesem Bereich sind der Hoffnung verpflichtet, mittels robuster und zuverlässiger Instrumente für die Analyse von Textoberflächen die mühevolle manuelle Wissensarbeit – etwa die Recherche, Formulierung und Modellierung beispielsweise von Definitionen für terminologische Datensammlungen – in Inhalt und Umfang zu reduzieren und zu erleichtern. Mit dem Verfügbarwerden umfangreicher elektronischer Textkorpora in den letzten Jahren und Jahrzehnten (als Beispiele für monolinguale Referenzkorpora seien hier das British National Corpus[1],das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts[2] – das allerdings Elemente eines elektronischen Wörterbuchs und eines Korpus‘ kombiniert –, das Russische Nationalkorpus[3] und das Corpus di Italiano Scritto[4] genannt, einen Überblick über weitere Korpora gibt Xiao 2008) und der Entwicklung von Methoden zur Konstruktion von Ad-hoc-Korpora mit Hilfe des Internets (vgl. Baroni/Bernardini 2004) haben entsprechende Arbeiten ein verstärktes wissenschaftliches Interesse erfahren: Die nunmehr bereit stehenden Textressourcen scheinen die ideale Grundlage für die Entwicklung semantischer Anwendungen darzustellen. Wesentlich für die Ausbeutung von Textkorpora als Wissensspeicher ist dabei eine Grundannahme der Korpuslinguistik, derzufolge bei ausreichender Korpusgröße Relevantes und Korrektes von Irrelevantem und Falschem klar unterschieden werden kann (vgl. Teubert/Čermáková 2007: 122):

> „The relevance principle of corpus linguistics assures that whatever is thought to be important will be repeated in other texts. Once our corpus is large enough to display a certain saturation of paraphrases, we can rest assured that what is missing is at least not the mainstream understanding of our unit of meaning.”

Hiermit ist schließlich auch der grundlegende methodologische Ansatz des bislang nur knapp skizzierten Forschungsfelds ins Spiel gekommen: die Analyse sprachlicher Regularitäten mit Hilfe *korpuslinguistischer* – quantitativer und symbolischer – Methoden, die dazu dienen sollen, relevante Wissensbestandteile zu identifizieren. Diese Vorgehensweise beruht auf der Annahme, dass semantisches Wissen, z. B. Wissen über semantische Relationen, die zwischen Begriffen bestehen, sprachlich durch wiederkehrende linguistische Strukturen, die zu sprachlichen Mustern und Regularitäten verallgemeinert werden können, vermittelt wird (vgl. Condamines 2002: 144):

> „The idea of associating lexical elements with a semantic relation is found under different names in different disciplines: linguistics, terminology, computer science (within knowledge acquisition from texts). To give a few linguistic examples, Lyons [...] was probably the first to evoke ’formulae’ used for fitting terms linked by a hyperonymic relation. Cruse [...] speaks of ‘diagnostic frames’.”

Gemeinsam ist den hiermit beschriebenen Ansätzen folglich, dass der Text – das in einem Korpus gesammelte sprachliche Material – als Ausgangspunkt für den Aufbau semantischer Strukturen betrachtet wird. Der Wissenserwerb aus Texten lässt sich entsprechend als *mapping* semantischer

[1] http://www.natcorp.ox.ac.uk/, Zugriff am 21.03.2014.
[2] http://www.dwds.de/, Zugriff am 21.03.2014.
[3] http://www.ruscorpora.ru/en/, Zugriff am 21.03.2014.
[4] http://sciling.dslo.unibo.it/coris_ita.html, Zugriff am 21.03.2014.

Strukturen auf sprachliche Elemente der Textoberfläche beschreiben, wobei statistische Verfahren häufig dann zum Einsatz kommen, wenn es darum geht, Relevantes von Irrelevantem zu trennen. Für die Terminologie bedeutet dies, dass klassische Methoden der Terminologiearbeit durch Techniken der linguistischen Korpusanalyse zu ergänzen sind. In der Terminologieforschung waren es insbesondere französischsprachige Forscher und Forscherinnen, die vor dem Hintergrund dieser Entwicklungen in anderen Disziplinen vehement für eine stärker automatisierte text- und sprachbasierte Terminographie unter Anwendung computerlinguistischer Technologien eintraten und die klassische, vom Begriff ausgehende Methode der Terminologiearbeit nach Wüster (1991[3]) ablehnten (z. B. Bourigault/Slodzian 1999, Van Campenhoudt 2001, L'Homme 2004, aber auch Pearson 1998)[5]. Die Chancen und Möglichkeiten einer textuell ausgerichteten und somit auf einem ausgesprochenen *bottom-up*-Ansatz beruhenden Terminologie für die praktische Terminologiearbeit und die Terminologiewissenschaft zu überprüfen, gehört zu den wesentlichen Anliegen dieser Arbeit.

1.2 Terminologiewissenschaftliche Perspektiven

Für die Terminologie bedeutet ein korpuslinguistischer Ansatz der terminographischen Beschreibung eine grundlegende Neuerung: Sind korpuslinguistische Techniken im Bereich der Termakquisition (automatische Termextraktion) mittlerweile akzeptiert (wenn auch nur in Teilen zufriedenstellend umgesetzt und in terminographische *workflows* integriert), fehlt für die auf den Erwerb terminologischer Einheiten folgenden Arbeitsschritte – hier insbesondere die begrifflich-semantische Beschreibung terminologischer Einheiten – noch weitgehend das Instrumentarium. Ansätze zur Nutzung korpuslinguistischer Techniken für die Terminographie existieren insbesondere für das Französische und das Englische (z. B. Biébow/Szulman 1999, Aussenac-Gilles/Jacques 2008, Barrière 2010). Allerdings haben all diese Ansätze einige wesentliche Nachteile gemeinsam:

- Sie sind in der Regel an bestimmte Sprachen gebunden, hier insbesondere Englisch und Französisch. Gemeinsam sind diesen Sprachen eine feste Wortfolge sowie schwach ausgeprägte Flexionsparadigmen – Eigenschaften, die auf eine Vielzahl terminologisch relevanter Sprachen nicht zutreffen. Eine sprachunabhängige Bewertung der angewandten Extraktionsmethoden ist daher kaum möglich. Zudem fehlt eine multilinguale Perspektive, die den Vergleich sprachlicher Oberflächenstrukturen für ähnliche semantische Kategorien und Relationen über Sprachgrenzen hinweg erlauben und mithin den Weg für elaboriertere Techniken der Wissensextraktion wie etwa die Alignierung wissenshaltiger Kontexte in multilingualen Textkorpora ermöglichen würde, fast vollständig.
- Sie sind häufig an ein bestimmtes Fachgebiet gebunden (z. B. *medical text mining*, vgl. Ben Abacha/Zweigenbaum 2011), domänenübergreifende Methoden wurden bislang nur in geringem Maße entwickelt.
- Es existiert kaum Wissen darüber, wie Menschen semantisches Wissen kategorisieren und speichern. Welche Formen und Arten semantischen Wissens für das Verständnis terminologischer Begriffe grundlegend sind, bleibt mithin unklar.

Obige Kritikpunkte machen deutlich, dass der gegenwärtige Forschungsstand in Bezug auf die Extraktion wissenshaltiger Kontexte aus Textkorpora weder den Erfordernissen der terminographischen Praxis noch den Bedürfnissen der Nutzer terminologischer Ressourcen gerecht wird. Vorliegende Arbeit strebt daher folgende Ziele an:

[5] L'Homme (2004: 17) prägt für eine so ausgerichtete Terminologie den Begriff „terminotique".

- Entwicklung, Implementierung und Evaluierung von Methoden für die Extraktion wissenshaltiger Kontexte für bislang nicht oder nur selten untersuchte, sich von den überwiegend bearbeiteten Sprachen Englisch und Französisch typologisch unterscheidende Sprachen. Konkret soll die Extraktion wissenshaltiger Kontexte für die Sprachen Deutsch und Russisch untersucht werden. Dies eröffnet nicht nur eine kontrastive Perspektive, sondern ermöglicht auch die Beurteilung existierender Extraktionsmethoden in Bezug auf ihre Leistungsfähigkeit unabhängig von einer bestimmten Sprache und leitet zu möglicherweise weiterführenden, multilingualen Arbeiten über.
- Die zu entwickelnden Methoden sollen für die Anwendung in unterschiedlichen Fachgebieten geeignet sein, d. h. es werden generische Methoden entwickelt. Für ihre Evaluierung werden daher Korpora verwendet, die unterschiedliche Fachgebiete abdecken. Der Nachteil einer solchen generischen Methode besteht freilich darin, dass sie den Besonderheiten stark spezialisierter Fachgebiete (z. B. Recht oder Medizin) unter Umständen nicht gerecht wird.
- Eine detaillierte Annotationsstudie und eine terminologieorientierte qualitative Bewertung von Extraktionsergebnissen sollen Aufschluss darüber geben, ob die Nutzer terminologischer Ressourcen – insbesondere Übersetzer – überhaupt einen Konsens in Bezug auf die Nützlichkeit bestimmter Aussagen entwickeln, und in welchem Maße automatisch extrahierte wissenshaltige Kontexte die terminologische Begriffsmodellierung unterstützen können.

Der hiermit umrissene Arbeitsansatz der semantischen Beschreibung terminologischer Begriffe anhand von Textkorpora – also auf Grundlage terminologischer Systematisierungen und Kategorisierungen durch die Nutzer der Terminologie selbst – entspricht mithin ebenfalls nicht der klassischen Terminologielehre nach Wüster (1991[3]), sondern versucht, sie durch Anwendung moderner linguistischer Methoden zu ergänzen und an eine empirische Faktenbasis zu binden. Angeknüpft wird dabei an eine Tradition der linguistischen Semantik, die empirisch-funktional vorgeht und semantische Bedeutungszuschreibungen aus beobachtbaren und quantifizierbaren Textmerkmalen abzuleiten versucht. Damit berührt das beschriebene Vorhaben eine Leitdifferenz innerhalb der modernen Linguistik (vgl. hierzu Geeraerts 2010), die sich auch in der terminologiewissenschaftlichen Fachdiskussion wiederfindet (vgl. Temmerman 2000), die Terminologiepraxis bislang jedoch kaum – positiv oder negativ – beeinflussen konnte. Weber (1999: 21, sic!) formuliert den allgemeineren Unterschied zwischen strukturalistischen *top-down*- und empirisch-funktionalistischen *bottom-up*-Ansätzen für die Linguistik wie folgt:

> „Die strukturale Definition spricht linguistischen Entitäten bestimmte morphologische, syntaktische und semantische Eigenschaften zu, aufgrund derer sie Kategorien wie Subjekt, Objekt, Prädikat etc. ausfüllen können. Die funktionale Sichtweise hingegen nimmt an, daß grammatische und semantische Kategorien auf diesen Relationen basieren und durch sie definiert werden."

Für die Terminologie folgt aus dem umrissenen korpuslinguistischen und empirischen Ansatz, dass begriffliche Systematisierungen nicht aus logischen *constraints* eines konkreten Begriffssystems oder den Ergebnissen einer strukturalistischen Komponentenanalyse abzuleiten sind, sondern durch textuelle Verknüpfungen des zu definierenden Begriffs (repräsentiert durch den entsprechenden Terminus) mit anderen Begriffen konstituiert werden. Beispielsweise charakterisieren Aussenac-Gilles et al. (2000: 174) ihre korpuslinguistisch fundierte Theorie der Terminologie wie folgt:

> „This linguistic theory is an interpretative, linguistic and textual semantics: its object is the attested text, the meanings are described through linguistic paraphrases, and the interpretation provides the

> meanings. Moreover, a concept is a normalized meaning, which results from a work of restriction with respect to the corpus and the application."

Die hier vorgetragene Konzeptualisierung des Begriffs als *normalized meaning* ist neuartig und stellt gegenüber radikal konstruktivistischen Ansätzen wie etwa dem von Bourigault und Slodzian (1999) insofern eine Verbesserung dar, als sie die Formulierung sprach- und repräsentativitätsbezogener Bewertungskriterien für Modellierungsentscheidungen erlaubt. Gegenüber der Wüsterschen Definition des Begriffs verspricht eine korpuslinguistische Herangehensweise, dass begriffliche Übereinstimmungen anhand textueller Okkurrenzen dargestellt werden können, die vormals mit den klassischen Mitteln der Terminographie nicht greifbare kognitive Größe Begriff also zumindest teilweise sprachlich operationalisierbar wird. Können zudem zwischenbegriffliche Relationen anhand sprachlicher bzw. textueller Kriterien verallgemeinert werden, lässt sich möglicherweise nicht nur die Beschreibung von Begriffstypen anhand finiter *sets* von Relationen, sondern auch die Strukturierung von Begriffssystemen anhand vorgegebener Kriterien auf einer empirischen Faktenbasis formalisieren. Helbig (2001: 8, sic!, Hervorhebungen im Original) hebt ebenfalls – allerdings für das allgemeinere Vorhaben der Wissensrepräsentation – die Vorteile einer Konzeptualisierung von Bedeutung als normalisierte Größe hervor, bezieht seine Ausführungen aber auf ganze Sätze:

> „Eine Frage, die oft im Zusammenhang mit der Wissensrepräsentation gestellt wird, ist die, ob eine **kanonische Bedeutungsdarstellung** existiert, d. h. ob sich eine Funktion angeben läßt, die semantisch äquivalente Sachverhalte in identische Bedeutungsstrukturen abbildet. [Es ist ...] nicht zu erwarten, daß es eine solche kanonische Bedeutungsrepräsentation überhaupt gibt, da bereits für wesentlich einfachere Strukturen als die natürliche Sprache keine solche Repräsentationsform existiert [...]. Was man allerdings erreichen kann, ist eine gewisse **Normalisierung** der Bedeutungsstrukturen natürlichsprachlicher Ausdrücke, indem man unter Verzicht auf bestimmte linguistische Feinheiten eine Einengung der Repräsentationsvielfalt durch Identifikation von Bedeutungsstrukturen für (fast) inhaltsgleiche Sätze [...] erzielt."

Wright (2007: 165) weist darauf hin, dass mit einer solchen korpus-basierten Normalisierung begrifflicher Information ein „[...] latent cognitive consensus reflected in texts as opposed to the practice of infividuals mapping their perceived structures [...]" abgebildet werde. Damit ist zugleich die Einschränkung formuliert, dass aus Korpora gewonnene Daten zwar keine absoluten Wahrheitswerte, wohl aber nach wissenschaftlichen Gesichtspunkten generierte und ausgewertete Datenagglomerationen darstellen können. Kann gezeigt werden – und dies gehört ebenfalls zu den Zielen vorliegender Arbeit –, dass eine korpusbasierte Vorgehensweise etablierte terminographische Techniken sinnvoll ergänzen kann, böte dies Gelegenheit für die Integration und Systematisierung alltäglicher Arbeitsmethoden von Terminologiepraktikern und Übersetzern (nämlich die Internetrecherche in Texten sowie allgemein die Verwendung von Paralleltexten) in die terminologische Theorie. Die Untersuchung wissenshaltiger Kontexte bietet für dieses im weitesten Sinne methodologische Projekt einen geeigneten Ansatzpunkt.

1.3 Bedeutung wissenshaltiger Kontexte für die terminologische Praxis

Wurden bisher die dieser Arbeit zugrunde liegenden theoretisch-methodischen Leitlinien im Hinblick auf ihren erhofften Beitrag zur aktuellen wissenschaftlichen Fachdiskussion herausgearbeitet, sollen nun auch terminologiepraktische Aspekte diskutiert werden. Einzugehen ist dabei zunächst auf eine deutlich wahrnehmbare Diskrepanz zwischen Theorie und Praxis: Fordern einschlägige Normen

(ISO 12620: 2009 in Nachfolge von Dubuc/Lauriston 1997) die Inklusion unterschiedlicher Kontexttypen in terminologische Ressourcen, erweist sich die fachliche Praxis in Gestalt existierender Ressourcen oft als weniger systematisch und meist sogar unvollständig. ISO-Normen für die terminographische Beschreibung von Termini und Begriffen sind etwa in den Standards ISO 12620: 2009 (ISO 2009a) sowie ISO 704: 2009 (ISO 2009b) und ISO 1087-1: 2000 (ISO 2000) definiert. Sieht ISO FDIS 12620: 1997 noch keinerlei kontextuelle Information im hier intendierten Sinn vor (vgl. Budin/Bühler 1999: 2105 f.), differenziert ISO 12620: 2009 unter anderem zwischen erklärenden und definitorischen Kontexten, die Eigenschaften von Begriffen und zwischen Begriffen bestehende Beziehungen explizieren.

Diese theoretischen Grundlagen kommen in der terminographischen Praxis jedoch häufig nicht zum Tragen. So gibt Keller (2006: 9 ff.) einen Überblick über eine große Bandbreite unterschiedlicher Ansätze zu Eintragsmodellen in Terminologieverwaltungssystemen, der zeigt, dass die Kategorie „Kontext“ je nach theoretischem Standpunkt des Autors entweder zu den benennungs- oder aber zu den begriffsbezogenen Daten, mitunter auch zu den pragmatischen Angaben gezählt, eine methodisch-theoretisch stringente Unterscheidung zwischen verschiedenen Kontexttypen und den durch sie vermittelten Informationsarten also nicht getroffen wird. Eine schlüssige Definition von Kontexttypen scheint in der Praxis mithin ebenso häufig zu fehlen wie eine klare Vorstellung dessen, was die terminographische Begriffsbeschreibung konkret zu leisten hat. Ein Vergleich einer Reihe von Einträgen großer, harmonisierter Terminologiedatenbanken wie IATE[6], TermiumPlus[7] und AkadTerm[8], aber auch von Online-Wörterbüchern wie Leo[9] und Multitran[10] und experimentellen terminologischen Ressourcen wie TermWiki[11] und EcoLexicon[12] (Stand März 2014) zeigt darüber hinaus, dass nur wenige Ressourcen um eine systematische Kontextrepräsentation bemüht sind: Einzig TermiumPlus verfügt über ein elaboriertes System der semantisch-linguistischen Annotation von Termini mit Hilfe von Definitionen und linguistischen sowie semantischen Kontexten. IATE, EcoLexicon und TermWiki enthalten für einige Termini unterschiedliche Typen von Definitionen.

Die wissenschaftliche Diskussion um die Bedeutung wissenshaltiger und anderer Kontexte – also Quellen semantischen und linguistischen Wissens – für die Terminologiearbeit wird indes seit mehr als 20 Jahren aktiv geführt, wobei die Wichtigkeit von Kontexten, die neben linguistischer auch semantische Informationen vermitteln, etwa im Zusammenhang mit dem Konzept der terminologischen Wissensbank (*terminological knowledge base)* deutlich hervorgehoben wurde. Der Begriff der terminologischen Wissensbank wurde 1992 von Meyer et al. geprägt, um der noch zu beschreibenden Unterscheidung zwischen unterschiedlichen Kontexttypen und der insbesondere für die Terminologie relevanten Differenzierung der durch diese Kontexttypen vermittelten Wissensarten (linguistisches vs. begriffliches Wissen, also *world knowledge,* Weltwissen oder Sach- bzw. enzyklopädisches Wissen, vgl. Engelberg/Lemnitzer 2008[3]: 7) Rechnung zu tragen (Meyer et al. 1992: 956, Hervorhebungen im Original):

> „[... W]hile terminology can certainly benefit from advances in computational lexicology, it nonetheless has its own non-trivial research problems, which are ultimately related to the quantity and types of *specialized world knowledge* that terminological repositories must contain. [...] A major weakness of

[6] http://www.iate.europa.eu, Zugriff am 21.03.2014.
[7] http://www.termiumplus.gc.ca, Zugriff am 21.03.2014.
[8] http://www.termini.lza.lv/term.php, Zugriff am 21.03.2014.
[9] www.leo.org, Zugriff am 21.03.2014.
[10] www.multitran.ru, Zugriff am 21.03.2014.
[11] www.termwiki.com, Zugriff am 21.03.2014.
[12] http://ecolexicon.ugr.es/en/index.htm, Zugriff am 21.03.2014.

> TDB[13]s is that they provide mainly *linguistic* information about terms (e.g. equivalents in other languages, morphological information, style labels); *conceptual* information is sparse (limited to definitions and sometimes contexts), unstructured, inconsistent and implicit."

Als Gegenentwurf zu einer rein lexikalisch ausgerichteten Präsentation terminologischer Daten entwickeln die Autoren das Konzept einer terminologischen Wissensbank, deren wichtigstes Charakteristikum die umfassende semantische Beschreibung der enthaltenen Begriffe ist (ebd.: 957, Hervorhebungen im Original): „[A] TKB must make explicit what a native speaker who is also a domain expert knows about concepts denoted by specialized lexical items". Eine terminologische Wissensbank umgeht folglich die für Lexikographie und Terminographie grundlegende Unterscheidung zwischen sprachlichem und enzyklopädischen Wissen und positioniert sich ähnlich wie die Gruppe der enzyklopädischen Wörterbücher (vgl. Engelberg/Lemnitzer 2008[3]: 14 ff.) auf einer zwischen diesen scheinbaren Antipoden gelegenen Mittelposition. Zwar ist die von Meyer et al. (1992) vorgetragene detaillierte Beschreibung terminologischer Einträge durch einschlägige ISO-Normen, insbesondere ISO 12620: 2009 (ISO 2009a), sowie realisierte Terminologiedatenbanken wie EuroTermBank[14] (Rirdance / Vasiljevs 2006) und Eintragsmodelle in Terminologieverwaltungssystemen (vgl. Keller 2006) längst überholt, die Forderung nach einer expliziten Darstellung umfassenden semantischen Wissens konnte in der praktischen Terminographie jedoch bislang nicht umgesetzt werden. Ähnlich wie Meyer et al. betont schließlich auch Helbig (2001: 19) für den Bereich der Wissensrepräsentation, die Beziehungen eines Begriffs zu anderen Begriffen seien zum Wesen des Begriffs zu zählen, und liefert so ein Argument für die Einbeziehung umfangreicher begriffsbezogener Informationen – insbesondere semantischer Relationen – in terminologische Nachschlagewerke und Datenbanken. Eine pointierte Begründung für die Notwendigkeit des Einbezugs relationaler Informationen in die Terminologiepräsentation beispielsweise in Form wissenshaltiger Kontexte gibt Gruntar Jermol in einer Studie zur Rechtsterminologie im Sprachenpaar Deutsch-Slowenisch (2010: 40):

> „Die meisten [...] zur Verfügung stehenden Hilfsmittel vermitteln entweder mehr oder weniger gelungene Übersetzungen und/oder Begriffsdefinitionen [...], wobei aber die Einsicht in semantische Verbindungen der Termini zueinander im Großen und Ganzen fehlt. Im Rahmen der Begriffsdefinition können zwar einige Beziehungen des Terminus zu anderen Begriffen angedeutet werden [...]. Dadurch wird aber bei Weitem nichts über die Stellung des jeweiligen Terminus im Rechtssystem ausgesagt, was für die Zuordnung zweier Termini [...] jedoch unerlässlich ist."

Problematisch an klassischen übersetzerischen Hilfsmitteln bzw. terminologischen Ressourcen sei, dass sie zwar „isolierte Termini und dadurch nur inhaltliche Bruchteile, nicht aber auch das (Fach)Wissen" vermittelten (ebd.):

> „Dies ist für Übersetzer jedoch oftmals unzureichend, insbesondere dann, wenn man nicht nur Sprachbarrieren überwinden muss, sondern auch Probleme zu bewältigen hat, die sich aus den Unterschieden in den Rechtssystemen von Ausgangs- und Zielkultur ergeben."

Die Frage, inwieweit neue Techniken der Terminologiearbeit, beispielsweise die Extraktion wissenshaltiger Kontexte aus Textkorpora dazu beitragen können, die genannten Schwächen traditioneller Ressourcen zu beheben, ist mithin an konkrete Erfordernisse der terminographischen Praxis geknüpft und eröffnet im Hinblick auf die Lösung existierender und bekannter Probleme neue

[13] Terminology data bases (Anmerkung der Autorin).
[14] http://www.eurotermbank.com/, Zugriff am 21.03.2014.

Perspektiven. Konkret sind von der in dieser Arbeit verfolgten Untersuchung wissenshaltiger Kontexte Informationen darüber zu erhoffen, inwieweit eine Systematisierung des für die terminographische Beschreibung von Begriffen relevanten Wissens überhaupt praktikabel und aufgrund der gewählten Daten umsetzbar ist. Mit der Beschreibung wissenshaltiger Kontexte für die Sprachen Deutsch und Russisch sowie der Entwicklung einer Extraktionsmethode auf Grundlage von Internet-Korpora werden zudem linguistische Informationen zur Oberflächengestalt von wissenshaltigen Kontexten in den genannten Sprachen erarbeitet sowie ein Ansatz zum automatisierten Erwerb solcher Kontexte erprobt und im Hinblick auf seine Anwendbarkeit für die terminographische Praxis evaluiert.

Eine weniger terminologiepraktische, sondern vielmehr logisch-philosophische und terminologietheoretische Begründung für die Beschäftigung mit wissenshaltigen Kontexten liefert Roche (2012). Im Unterschied zu Vertretern einer korpusbasierten Terminologiearbeit knüpft er explizit an die Terminologietheorie Wüsters (1991[3]) an. Roche (2012) unterscheidet dabei zunächst zwischen essentiellen und deskriptiven Eigenschaften (Attributen) eines Begriffs. Semantische Relationen, die in der Terminologielehre zur Beschreibung eines Begriffs akzeptiert sind, etwa partitive und assoziative Relationen, bezeichnet Roche als deskriptiv (ebd.: 24) und zitiert im Sinne einer Ergänzung klassischer terminologischer Begriffsdefinitionen durch deskriptive Elemente aus der historischen Arbeit „Logik oder die Kunst des Denkens“[15] (Arnauld/Nicole 1996, zitiert nach ebd.: 27):

> „There are two kinds of definitions: the more exact one, which retains the name definition, and the other, less exact, which is called a description. The more exact definition explains the nature of a thing by its essential attributes, of which the common one is called the genus, and the proper one the difference. The less exact definition, called a description, provides some knowledge of a thing in terms of the accidents that are proper to it and determine it enough to give us an idea distinguishing it from other things.”

Obwohl in vorliegender Arbeit vorrangig aus korpuslinguistischer Sicht argumentiert werden soll, kann somit die Beschäftigung mit wissenshaltigen Kontexten als Beispielen für Begriffsdeskriptionen im Sinne einer Ergänzung klassischer Definitionen durch weitere Informationsbestandteile bzw. einer überhaupt erst zu leistenden grundlegenden und ungefähren Konzeptbeschreibung auch aus klassisch-wüsterscher Sicht motiviert werden. Zudem ist anzumerken, dass im Kontext neuerer Forschungen zu semantischen Netzen semantische Beziehungen die Grundlage moderner wissensbasierter Anwendungen sind und die Implementierung einer Vielzahl auch für die Terminologie relevanter Funktionalitäten (vgl. Reichenberger 2010) ermöglichen. Diese Möglichkeiten hat auch Meyer (2001: 299, Hervorhebungen im Original) im Blick, wenn sie schreibt:

> „In principle, the [...] concept of knowledge patterns[16] also allows for a conceptual, *relation*-oriented approach, i. e., with the terminographer asking the system to show all term-sets that are linked by relation X. For example, the HYPERONYMY relation would generate a skeletal generic-specific hierarchy, probably an excellent starting-point for a terminographer just beginning to explore a new domain.”

[15] Hierbei scheint es sich um die 1662 im französischen Original erschienene Arbeit „La logique ou l’art de penser“ von Antoine Arnauld zu handeln, vgl. http://d-nb.info/gnd/4395825-4/about/html (Zugriff am 21.03.2014). In der Bibliographie ist die Arbeit in der von Roche (2012) zitierten englischen Übersetzung von 1996 aufgeführt.

[16] Hiermit sind linguistische Muster gemeint, mit deren Hilfe wissenshaltige Kontexte in Texten identifiziert werden können, vgl. Kapitel 3.

An ähnliche Szenarien scheinen auch Boguraev und Pustejovsky (1996: 9, zitiert nach Kunze/Lemnitzer 2007: 238) zu denken:

> „[A lexical knowledge base is] ... a large-scale repository of lexical information, which incorporates more than just static descriptions of words [...]. It is essentially a dynamic object, as it incorporates, in addition to its information types, the ability to perform inferences over them and thus induce word meaning in context."

Die Integration wissenshaltiger Kontexte in Terminologiedatenbanken würde vor dem Hintergrund dieser Überlegungen völlig neue Perspektiven der Terminologiearbeit eröffnen, da wissenshaltige Kontexte auf den Erwerb semantischen Wissens – also der Wissenskomponente der vorstehend zitierten Ansätze – fokussiert sind. Anzumerken ist hier allerdings, dass vorliegende Arbeit sich auf die Beschreibung und Extraktion wissenshaltiger Kontexte aus Korpora beschränkt – die Implementierung semantischer Anwendungen auf Grundlage von wissenshaltigen Kontexten ist ausdrücklich nicht Gegenstand der hier beschriebenen Forschung. Dennoch kann an dieser Stelle die Hypothese festgehalten werden, dass wissenshaltige Kontexte einen wesentlichen Schritt hin zu einer stärker wissensbasierten und den Bedürfnissen von Übersetzern und Dolmetschern besser gerecht werdenden Terminologiearbeit bedeuten. Vorliegende Studie möchte einen Beitrag zu einer terminologiepraktisch ebenso wie methodologisch-theoretisch orientierten Überprüfung dieser Hypothese leisten.

1.4 Vorgehensweise und Gliederung

Abschließend soll nun ein Überblick über die dieser Arbeit zugrunde liegende eigene Vorgehensweise gegeben werden. Diese ist durch die folgenden Aspekte charakterisiert:

- Nutzung des Internets als Datenquelle, insbesondere in Form von *web*-Korpora,
- paralleles zweisprachiges Vorgehen,
- iterative Analyse sprachlicher Eigenheiten wissenshaltiger Kontexte, Implementierung und Evaluierung einer *baseline*-Methode und sukzessive Verbesserung der Methode.
- Evaluierung von Extraktionsergebnissen anhand eines Goldstandards sowie mit Hilfe von Entwicklungs- und Testkorpora.

Vorliegende Arbeit nutzt das Internet als prominente Datenquelle und hebt sich damit von klassischen korpuslinguistischen Studien ab. Begründen lässt sich dieser Umstand damit, dass Übersetzer und Dolmetscher – und damit die Hauptnutzer terminologischer Informationsangebote – das Internet bereits seit Jahren zu Zwecken der Terminologierecherche nutzen. Sie tragen damit zwei wesentlichen sachlichen Zwängen der übersetzungsbezogenen Terminologiearbeit Rechnung, und zwar einerseits der Knappheit relevanter Informationen im Angesicht der Vielzahl der von ihnen bedienten Sprach- und Fachgebietsgruppen sowie andererseits einem üblicherweise hohen Zeitdruck. Die Nutzung von *web*-Korpora steht dabei im Vergleich zu Studien auf Grundlage klassischer Korpusmaterialien (z. B. wissenschaftlicher Artikel) vor zusätzlichen Herausforderungen – beispielsweise dem Umgang mit sprachlichen oder inhaltlichen Fehlern, Dopplungen, etc. –, muss aber insofern als gerechtfertigt gelten, als eine für die praktische Terminologiearbeit nutzbare Methode Ansatzpunkte für die Arbeit mit relevanten Sprachen und Fachgebieten über die bereits wohluntersuchten und mit Ressourcen bestens ausgestatteten Bereiche (für die Sprachen insbesondere das Englische, für die Fachgebiete insbesondere die Medizin) hinaus bieten muss. Durch die Untersuchung unterschiedlicher Fachgebiete anhand deutscher und russischer Texte aus

dem Internet wird diese Arbeit dem soeben formulierten Anspruch gerecht. Die Nutzung von Internetquellen scheint darüber hinaus auch im Hinblick auf die zunehmende Nutzung der online vorfindbaren Datenfülle für komplexe Studien im Bereich der Informationsextraktion gerechtfertigt: Sollen solche Forschungsansätze für die Terminologie nutzbar gemacht und auf neue Sprachen übertragen werden (sowohl Xu (2007) als auch Blohm (2010), um nur zwei Beispiele zu nennen, beschäftigen sich ausschließlich mit englischen Texten), ist zu eruieren, welche Art von Informationen aus „schmutzigen" *web*-Korpora überhaupt zu extrahieren sind und mit welcher Qualität.

Diese Studie wählt eine iterative und parallele Vorgehensweise. In **Kapitel 2** wird zunächst der Begriff des wissenshaltigen Kontexts theoriebasiert geschärft und von ähnlichen Begriffen abgegrenzt. **Kapitel 3** umreißt auf Grundlage einer Darstellung des relevanten Forschungsstandes den theoretisch-methodologischen Rahmen, innerhalb dessen sich die empirische Arbeit im Zusammenhang mit dieser Studie bewegt. Sowohl für das Deutsche als auch das Russische sollen alsdann parallel Textkorpora aufgebaut werden, in denen wissenshaltige Kontexte zur Erstellung eines Goldstandards manuell identifiziert und bewertet werden (**Kapitel 4**). Auf Grundlage dieser Daten soll im weiteren Verlauf der Arbeit eine linguistische Analyse Hinweise auf die spezifischen Eigenschaften wissenshaltiger Kontexte erbringen (**Kapitel 5**). Die parallele Untersuchung russischer und deutscher Texte kann dabei zusätzlich Hinweise auf sprachliche Unterschiede zwischen relevanten Kontexten bzw. sprachübergreifende Gemeinsamkeiten geben. Im Ergebnis dieser Untersuchungen werden symbolische, d. h. musterbasierte Extraktionsverfahren entwickelt, implementiert, evaluiert und unter Zuhilfenahme statistischer Methoden verbessert (**Kapitel 6**). Die Evaluierung erfolgt dabei für jede Sprache anhand sogenannter Entwicklungs- und Testkorpora. Dies erlaubt die Überprüfung der Performanz sprachlicher Muster, die auf Grundlage der Entwicklungskorpora erarbeitet wurden, in einem neuen Extraktionssetting und gibt damit Hinweise auf deren Allgemeingültigkeit. Aus diesem Grund wird darauf geachtet werden, dass die Testkorpora auch vollständig neue Domänen umfassen. **Kapitel 7** schließlich fasst die erarbeiteten Ergebnisse zusammen und formuliert Desiderata für zukünftige Forschungsarbeiten.

Der Vollständigkeit halber ist noch eine methodologische Anmerkung anzubringen: Vorliegende Arbeit folgt einem korpuslinguistischen Paradigma und setzt im Zusammenhang mit ebendiesem eine Vielzahl automatischer Werkzeuge ein. Im Hinblick auf eine solche Vorgehensweise merken Neumann und Hansen-Schirra (2012: 23) an:

> „Quantitative research that includes linguistic enrichment instead of working with raw data as is done in the corpus-driven research paradigm […] depends on automatic annotation since the amount of text involved […] cannot be processed manually, particularly if the annotation is to comprise several layers. The annotation can therefore only be as accurate as the tools used […] ."

Es handelt sich hierbei um einen relevanten Hinweis: Bei der Verwendung von Methoden der automatischen Sprachverarbeitung ist mit Verarbeitungsfehlern und – für das hier verfolgte Vorhaben der automatischen Extraktion wissenshaltiger Kontexte besonders kritisch – mit der Propagation von Verarbeitungsfehlern auf nachfolgende Annotationsstufen zu rechnen. Dieser im Rahmen der vorliegenden Arbeit nur in geringem Maße zu beeinflussende Umstand ist sowohl bei der Bewertung von Extraktionsergebnissen als auch bei der Interpretation linguistischer Analysen stets zu berücksichtigen.

2. Wissenshaltige Kontexte

2.1 Terminologisch-methodologische Vorbemerkung

In diesem Kapitel soll eine theoriebasierte Definition wissenshaltiger Kontexte erarbeitet werden, welche die Grundlage für die spätere Annotation, Beschreibung und Extraktion russischer und deutscher Kontexte aus Textkorpora bilden soll. Als Terminus für die Bezugnahme auf den hier interessierenden Untersuchungsgegenstand wird dabei in Anlehnung an Meyer (2001) die Benennung „wissenshaltiger Kontext“ gewählt, obwohl in der relevanten wissenschaftlichen Literatur unterschiedliche Benennungen auftauchen. Gegen die Verwendung der deutschen Benennung „Mikrokontext“ (Hohnhold 1999: 2157) als Alternativvariante spricht dabei, dass Hohnhold als Mikrokontexte zwar lokale Kontexte von Termini beschreibt, die semantisch-informative Funktion wissenshaltiger Kontexte mit der linguistisch-informativen Funktion anderer Kontexttypen dabei aber vermischt:

> „Kein anderer Anwenderbereich von Terminologie bezieht so ausdrücklich das kontextuelle Umfeld der Termini in die Arbeit mit ein wie Übersetzer. Die Mitaufzeichnung von Mikrokontexten (zusammen mit dem jeweiligen Terminus) ist konstitutiver Bestandteil der ÜoT[17] [...]. Mikrokontexte liefern dem Übersetzer entweder phraseologisches Anschauungsmaterial [...] und/oder definitorische Anhaltspunkte bei fraglichen Begriffen.“

Das Zitat Hohnholds benennt indes die bereits einleitend hervorgehobene wesentliche Funktion wissenshaltiger Kontexte und bezeichnet diese als zentralen Bestandteil der übersetzungsorientierten Terminologiearbeit. ISO 12620:2009[18] definiert den Begriff *context* ganz ähnlich wie Hohnhold als „[...] text which illustrates a concept or the use of a designation“. Um die Trennung zwischen semantisch informativen und linguistisch informativen Kontexten auch terminologisch zu verdeutlichen, soll jedoch auch dieser Terminus zugunsten der Benennung „wissenshaltiger Kontext“ zurückgestellt werden. In den folgenden Abschnitten wird der Begriff des wissenshaltigen Kontexts weiter geschärft.

Die Beschreibung wissenshaltiger Kontexte ist eng verwandt mit der Beschreibung von Begriffsdefinitionen in Texten. Anders als für die Beschreibung wissenshaltiger Kontexte liegt jedoch für die Beschreibung und Klassifizierung von Definitionen bereits ein umfangreiches Korpus wissenschaftlicher Literatur vor. Um vorliegende Arbeit möglichst konzis und präzise zu halten, sollen die dort vorgetragenen Ansätze an dieser Stelle nicht wiederholt werden, vielmehr soll das Hauptaugenmerk dieses Kapitels auf die begrifflich-theoretische Schärfung des Untersuchungsgegenstands „wissenshaltige Kontexte“ sowie dessen Abgrenzung von anderen Kontexttypen gelegt werden. Für Diskussionen zur linguistischen Form und theoriegebundenen Klassifikation von Begriffsdefinitionen sei auf die Arbeiten Longworths (2006), Hanks‘ (2006b), Cramers (2011), Stanaitytės (2005, die beiden letztgenannten Arbeiten beschäftigen sich mit deutschen Alltagsdefinitionen) sowie für eine Diskussion der Besonderheiten von

[17] Übersetzungsorientierte Terminographie, Anmerkung der Autorin.
[18] Zitiert nach ISOCat, Zugriff am 25.01.2011.

Wörterbuchdefinitionen auf die Arbeiten Wiegands (1989), Alshawis (1989) sowie Vossens et al. (1989, die beiden letzten Publikationen bieten insbesondere eine computerlinguistische Perspektive) und Barnbrooks' (2002) verwiesen. Relevante Literaturüberblicke finden sich auch in Pearson (1998: 89 ff.) und Walter (2010: 53 ff.). Eine Diskussion mit Übersetzungsbezug bietet Oberholzer (2002).

2.2 Kontexttypen in der Linguistik und erste Beschreibung sowie Abgrenzung wissenshaltiger Kontexte von anderen Kontexttypen

Der linguistische Begriff des „Kontexts" ist überaus vielschichtig und bezieht sich häufig nicht auf linguistische Elemente, sondern vielmehr auf psychologische, pragmatische und kulturelle Gesichtspunkte der Kommunikation. So nennt Hanks (2006a: 115 f.) handlungsbezogene, soziale und psychologische Aspekte als maßgebliche Kontextelemente. Bußmann (2002[3]: 374) differenziert dagegen zwischen drei elementaren Kontexttypen, und zwar:

- dem generellen Situationskontext, in dem Kommunikation stattfindet (einschließlich zeitlicher und räumlicher Parameter),
- dem persönlichen und sozialen Kontext, der die Beziehung zwischen den Kommunikationspartnern beschreibt,
- dem linguistischen Kontext einer Äußerung.

Der Begriff des Kontexts steht zudem mit dem der Relation in enger Beziehung. Bereits de Saussure (1966: 122 ff.) unterscheidet syntagmatische und assoziative Relationen lexikalischer Einheiten, die den linguistischen Kontext dieser Einheiten definieren. Während erstere auf der linearen Natur sprachlicher Äußerungen beruhten, seien letztere individuell-kognitiv determiniert, d. h. an linguistische Einheiten gebundene Bewusstseinsinhalte, und grundsätzlich nicht vorhersehbar. Syntagmatische Relationen lassen sich daher dem sprachlichen Kontext nach Bußmann zuordnen, wohingegen assoziative Relationen im Sinne de Saussures auf der zweiten Ebene des persönlichen und sozialen Kontexts zu verorten sind.

Sollen nun wissenshaltige Kontexte einem der drei genannten Kontexttypen zugeordnet werden, ist insbesondere zu bedenken, dass wissenshaltige Kontexte hybride Gebilde, also Gruppen sprachlicher Einheiten darstellen, die auf das durch eine Fach-Community geteilte nichtsprachliche Wissen verweisen. Formal zu verorten sind wissenshaltige Kontexte mithin auf der Ebene des linguistischen Kontexts oder Ko-Texts (Sebeok 1986: 151) eines Terminus', d. h. innerhalb seiner syntagmatisch-syntaktischen Umgebung – eine Festlegung, welche die Abgrenzung wissenshaltiger Kontexte von nichtlinguistischen Kontexttypen wie etwa den von Hanks genannten Aspekten ermöglicht. Allerdings ist auch die weitere Einschränkung zu beachten, dass als wissenshaltige Kontexte nur diejenigen Ko-Texte eines Terminus' gelten können, die durch ihre sprachliche Oberflächenstruktur auf begrifflich-semantische Tiefenstrukturen verweisen[19]: Wissenshaltige Kontexte manifestieren sich innerhalb der unmittelbaren linguistischen Umgebung eines Terminus', explizieren dabei aber begrifflich-semantische Zusammenhänge. Mithin können wissenshaltige Kontexte hinsichtlich der durch sie vermittelten Informationen auch von anderen für die Terminologiearbeit relevanten

[19] Zur Unterscheidung zwischen Oberflächen- und Tiefenstruktur vgl. Helbig 2001: 15. Im Rahmen der hier vertretenen korpuslinguistischen Argumentation wird die Tiefenstruktur, wie einleitend dargestellt, als *normalized meaning* konzeptualisiert.

Kontexttypen wie Kollokationen (Picht 1987, vgl. auch Herbst/Klotz 2003: 83 ff.), lexikalischen Relationen (Synonyme, Antonyme, verwandte Wörter) und lexikalischen Funktionen (vgl. Mel'čuk et al. 1995) sowie linguistisch-illustrativen Verwendungsbeispielen abgegrenzt werden: Während linguistische Kontexttypen wie die soeben genannten vor allem linguistische Informationen zum korrekten Gebrauch der Benennung vermitteln, sollen wissenshaltige Kontexte ein besseres Verständnis des fraglichen Begriffs ermöglichen – sie vermitteln nicht linguistisches, sondern begriffsbezogenes Wissen. Assoziativ-individuelle und rein sprachliche Aspekte werden durch wissenshaltige Kontexte somit nicht erfasst. Abbildung 1 gibt einen graphischen Überblick über die diskutierten Aspekte linguistischer, wissenshaltiger und kommunikativer Kontexte:

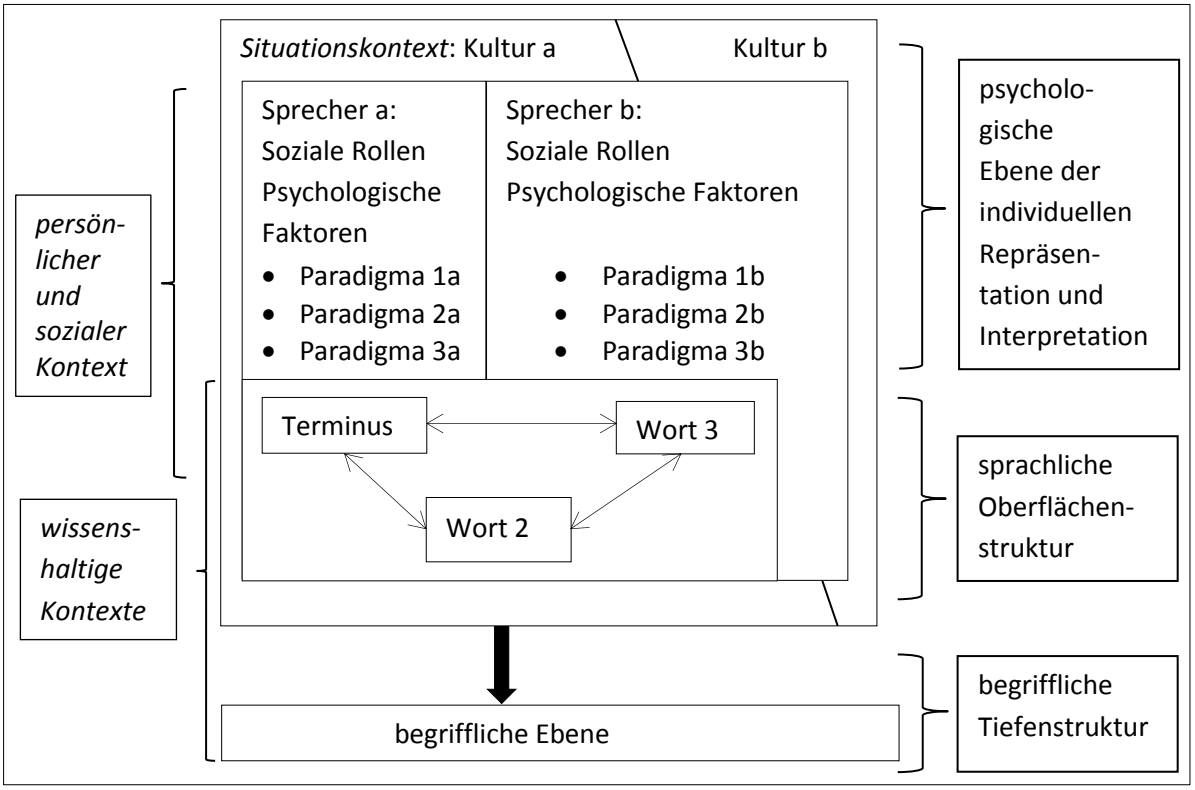


Abbildung 1: Kontexttypen und wissenshaltige Kontexte

Die größere Ebene im Hintergrund der Abbildung visualisiert den Situationskontext (einschließlich kultureller und individueller Prägungen), der den Gesprächspartnern zu gewissen Teilen gemeinsam, aber nicht notwendigerweise zwischen ihnen identisch ist und unterschiedliche Aspekte des Gesprächsthemas beeinflussen kann (diagonale Trennlinie zwischen Kultur a und Kultur b). Im Rahmen dieses Kontexts sowie vor dem Hintergrund sozialer Rollen und psychologischer Faktoren greifen Sprecher auf individuelle Wissensrepräsentationen zu (Paradigmen 1a bis 3b) und interpretieren Äußerungen in individueller Weise (mittlere Ebene, persönlicher und sozialer Kontext), wobei die mit den lexikalischen Einheiten der inneren Ebene (Terminus, Wort 2 und Wort 3, sprachliche Oberflächenstruktur), d. h. der Ebene der konkreten sprachlichen Äußerungen, verknüpften individuellen assoziativen Paradigmen der mittleren Ebene, d. h. Bewusstseinsinhalte bzw. Repräsentationen, die zwar an die lexikalischen Einheiten gebunden sind, aber nicht mit den etwa in Wörterbüchern kodifizierten Bedeutungen und auch nicht mit den individuellen Paradigmen anderer Personen sowie insbesondere nicht mit den durch wissenshaltige Kontexte versprachlichten überindividuellen Strukturen der begrifflichen Ebene übereinstimmen müssen, zum Tragen kommen

(assoziative Relationen nach de Saussure). Merkmale sowohl der äußeren als auch der mittleren Ebene können die sprachliche Gestalt von Äußerungen auf der inneren Ebene auf vielfältige Art und Weise beeinflussen, sind jedoch in erster Linie für das situationsbezogene Verstehen von Äußerungen und das angemessene Reagieren auf sie von Bedeutung. Aus semantischer Sicht beziehen sich die mittlere und äußere Ebene mithin auf die soziale und emotionale Bedeutung von Äußerungen (Wörtern und Sätzen) und sind daher von der durch wissenshaltige Kontexte vermittelten, eher als deskriptiv zu klassifizierenden Bedeutung (vgl. Löbner 2003: 23 ff.) als einer – dies ist bereits bei Wüster (1991[3]) entsprechend angelegt – überindividuellen und verallgemeinernden Wissensbasis für die Kommunikation in spezialisierten Wissensbereichen nicht nur formal (wissenshaltige Kontexte als unmittelbarer Ko-Text eines Terminus‘ im Gegensatz zu anderen, eher auf den Gesamttext bezogenen Kontextaspekten der mittleren und äußeren Ebene), sondern auch inhaltlich und funktional abzugrenzen.
Die innerste Ebene der lexikalischen Einheiten stellt in Abbildung 1 das textuelle Gewebe (Termini und Ko-Texte) dar. Die Einschränkung auf lediglich drei sprachliche Elemente ist hier aus darstellungstechnischen Gründen vorgenommen worden und rein symbolischer Natur. Innerhalb des Ko-Texts wiederum ist es allein die Einheit aus einem Terminus und seiner unmittelbaren syntaktisch-syntagmatischen Umgebung (syntagmatische Relationen nach de Saussure), welche in Gestalt wissenshaltiger Kontexte möglicherweise die Analyse begrifflicher Tiefenrelationen anhand syntagmatisch-syntaktischer und anderer Oberflächenphänomene ermöglicht. Diese Fassung wissenshaltiger Kontexte schließt damit auch erklärende und definierende Textstellen, die sich über mehrere Sätze hinweg erstrecken und mithin auf diskursiver Ebene untersucht werden müssen, aus dem Untersuchungsgegenstand vorliegender Arbeit aus.

2.3 Wissenshaltige Kontexte: Begriffsdefinition und Klassifikation

2.3.1 Inhaltliche Definition

Der Begriff des wissenshaltigen Kontexts (*knowledge-rich context*, KRC) wurde von Meyer (2001) geprägt, Ansätze zur Beschreibung und automatischen Extraktion wissenshaltiger Kontexte für die Terminographie finden sich jedoch bereits in früheren Arbeiten (vgl. Condamines/Rebeyrolle 1998). Meyer (2001: 280) stellt unter Bezugnahme auf Sowa (1984) fest, dass begriffliches Wissen logisch in zwei Kategorien zerfällt, nämlich Attribute und Relationen. Attribute machen dabei das intrinsische Wesen eines Begriffs aus, ohne den Begriff auf andere Begriffe zu beziehen – es handelt sich um die invarianten intrinsischen Merkmale von Begriffen, die es aus strukturalistischer Sicht zu bestimmen gilt. Damit stimmen Sowas logische Attribute im Wesentlichen mit Wüsters (1991[3]: 16) begrifflichen „Eigenmerkmalen“ überein. Relationen dagegen verbinden einen Begriff mit anderen Begriffen und sind nach Helbig (2001: 19) ebenfalls ein wesentlicher Bestandteil des Begriffsgehalts. Ein wissenshaltiger Kontext kann demzufolge als Textstelle, die Auskunft über ein relevantes Attribut oder eine relevante Relation eines bestimmten Begriffs gibt, definiert werden (Meyer 2001: 281, Hervorhebung im Original):

> „By *knowledge-rich context*, we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation.”

Ein solcher Kontext, so Meyer (ebd.: 282), liefert entweder bereits eine gültige Begriffsdefinition bzw. erste Anhaltspunkte für die Formulierung einer solchen. Über die inhaltliche Minimaldefinition Meyers hinaus sollte – zur Absicherung einer kohärenten Evaluation späterer Extraktionsergebnisse – die Beschreibung wissenshaltiger Kontexte an ein semantisches Modell zur Beschreibung von Sachverhalten geknüpft werden, aus dem wesentliche inhaltliche Bestandteile wissenshaltiger Kontexte abgeleitet werden können. Ein solches Modell beschreibt Xu (2007) für das Gebiet der Informationsextraktion. Gegenüber Meyers Minimalanforderungen führt Xu (ebd.: 49 f.) die zusätzlichen Elemente „zeitliche Anbindung“ (einer semantischen Relation) und „Ereignis“ an, wobei letzteres für die Beschreibung wissenshaltiger Kontexte – anders als für die Informationsextraktion – allerdings insofern von untergeordneter Bedeutung scheint, als die Terminographie weniger an der Beschreibung punktueller Ereignisse als am Aufzeigen stabiler Wissenszusammenhänge interessiert ist (Barrière 2004: 8 ff). Auch die von Xu angeführten komplexen Relationen – Beziehungen zwischen Begriffen und Relationen – sollen vorerst nicht betrachtet werden, da für die terminologische Beschreibung von Begriffen einfache, grundlegende Relationen bedeutsamer scheinen als komplexe Beziehungen. Walter und Pinkal (2006) nennen in ihrer Studie zur Extraktion von Definitionen aus deutschen Gerichtsurteilen zudem die Benennung eines Verwendungsgebiets als wesentliches semantisches Beschreibungselement. Abbildung 2 visualisiert das hieraus entstehende semantisch-sprachliche Modell wissenshaltiger Kontexte:

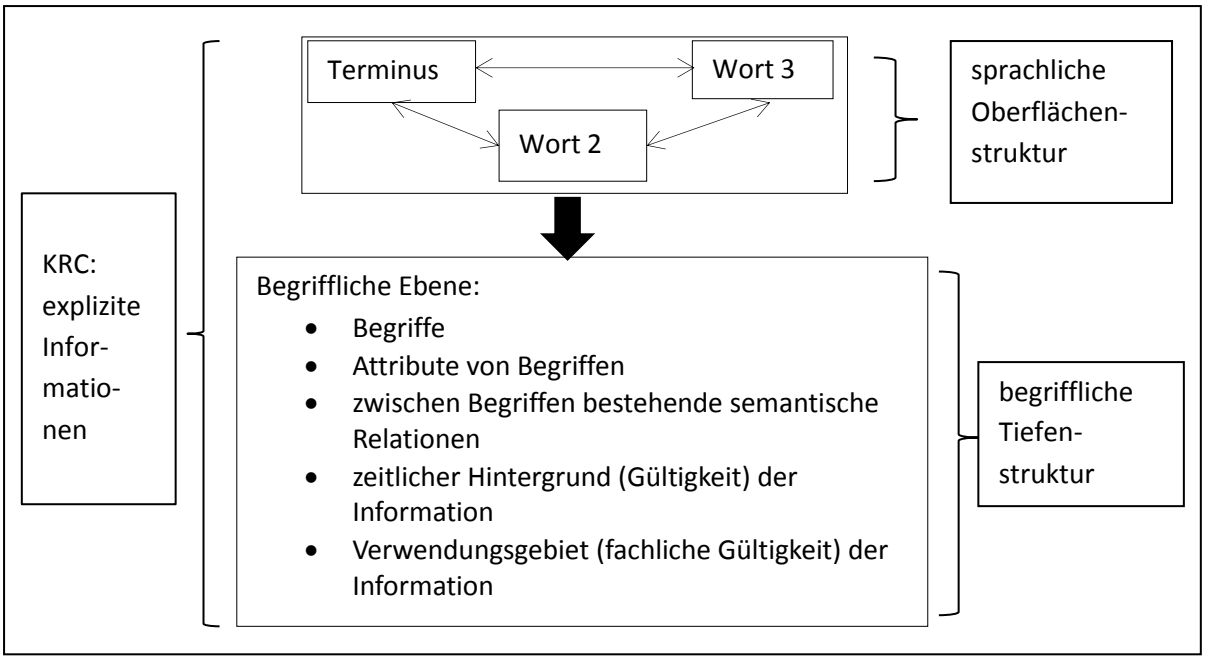


Abbildung 2: Semantisch-sprachliches Modell wissenshaltiger Kontexte

Festzuhalten ist zudem, dass wissenshaltige Kontexte nicht isoliert in Lexika und Wörterbüchern auftreten, sondern aus dem Zusammenhang größerer Texte und Textsammlungen entnommen werden können. Anders als im Fall von Definitionen ist an sie nicht der Anspruch einer vollständigen und umfassenden Begriffsklärung geknüpft, vielmehr erfüllen wissenshaltige Kontexte situationsspezifische Kommunikationserfordernisse und beleuchten unterschiedliche Facetten eines Begriffs. Eine inhaltliche Prüfung wissenshaltiger Kontexte bzw. ein spezifischer Umgang mit ihrem semantischen Gehalt ist für die Terminologie im Hinblick auf Gesichtspunkte der Qualitätssicherung mithin unumgänglich. Als Weiterentwicklung der Minimaldefinition von Meyer (2001) kann nun eine eigene Definition wissenshaltiger Kontexte formuliert werden:

> Unter wissenshaltigen Kontexten werden in Texten und Textsammlungen natürlich vorkommende Textbestandteile (Sätze) verstanden, die in einer für das Verständnis des Begriffs durch den Leser möglicherweise förderlichen Weise Attribute fachspezifischer Begriffe und zwischen ihnen bestehende semantische Relationen vor einem relevanten zeitlichen Hintergrund explizit darstellen.

Diese Definition akzentuiert gegenüber derjenigen Meyers einige zusätzliche Aspekte:

- Betont wird das natürliche Vorkommen wissenshaltiger Kontexte. KRCs sind mithin aus Gebrauchstexten unterschiedlichster Herkunft zu extrahieren, sollen sie doch Aufschluss darüber geben, welche begrifflichen Strukturen durch die Nutzer der Terminologie selbst verwendet werden. Wörterbuchdefinitionen sind mithin nicht als KRCs zu betrachten.
- Die in einem KRC enthaltene Information sollte das Verständnis eines Begriffs durch den Leser fördern. Die Nachweisführung, dass dies mit Hilfe automatisch aus einem Textkorpus extrahierter Kontextkandidaten gelingen kann, könnte Inhalt einer eigenen Untersuchung sein.
- Die Information muss im Sinne der einleitend zitierten Ansätze zu terminologischen Wissensbanken explizit dargestellt sein.
- Definitionen werden als Sonderfälle wissenshaltiger Kontexte aufgefasst, die sich durch einen der klassischen intensionalen (aristotelischen) Definitionsstruktur entsprechenden Aufbau auszeichnen.
- Implizit legt die formulierte Definition auch fest, dass die in einem KRC enthaltene Information möglichst generalisierbar, also allgemeingültig, sein sollte. Das Merkmal der Generalisierbarkeit wird beispielsweise auch von Walter (2010: 52) als wesentliches Definitionsmerkmal genannt.

2.3.2 Klassifikation wissenshaltiger Kontexte

2.3.2.1 Formale Klassifikation

Ansatzpunkte für eine formale Klassifikation wissenshaltiger Kontexte liefert die Unterscheidung zwischen definitorischen, erklärenden und assoziativen Kontexten, die in der Terminologie bereits getroffen wurde, bevor wissenshaltige Kontexte auch computerlinguistisch systematisch erforscht wurden. ISO 12620: 2009 (ISO 2009) beschreibt Kontexte für die Terminographie und nennt neben übersetzten, linguistischen und metalinguistischen Kontexten, die hier als Elemente anderer Kategorisierungsebenen außer Acht gelassen werden sollen, auch assoziative, erklärende und definitorische Kontexte. Damit folgt die ISO-Norm weitgehend der Beschreibung dieser Kontexttypen durch Dubuc und Lauriston (1997). So wird ein definitorischer Kontext als Textstelle beschrieben, die zwar umfassende Informationen über einen Begriff enthält, aber nicht die formale Strenge einer Definition besitzt (ISO 12620: 2009, zitiert nach ISOCAT[20]). Ein erklärender Kontext enthält dagegen lediglich eine zusammenfassende Erklärung des Begriffs und ein assoziativer Kontext ermöglicht schließlich nur noch die Zuordnung eines Begriffs zu einem bestimmten Fachgebiet oder aber einer speziellen Wortform (ebd.).
Wird diese Klassifizierung auf wissenshaltige Kontexte übertragen, ist in Anknüpfung an die bereits in der Einleitung zitierte Arbeit von Roche (2012, vgl. auch Cramer 2011: 85) indes festzustellen, dass

[20] http://www.isocat.org/interface/index.html, Zugriff am 25.08.2011.

wissenshaltige Kontexte zwar durchaus die Form klassischer Definitionen aufweisen können, aus epistemologischer Sicht aber nur durch zusätzliche Prüfung – etwa in Form klassischer Terminologiearbeit – festgestellt werden kann, ob die durch den Kontext dargestellten Begriffseigenschaften tatsächlich essentieller oder aber deskriptiver Natur sind. Die Unterscheidung zwischen definitorischen und erklärenden Kontexten scheint mithin für die Praxis nur bedingt relevant zu sein. In der Tat ist die Unterteilung von Kontexten in definitorische, erklärende und assoziative Typen daher eher als formal-theoretische Klassifikation zu betrachten.

2.3.2.2 Epistemologische Klassifikation

Am Ende des vorigen Abschnitts war bereits angedeutet worden, dass eine rein formale Klassifikation wissenshaltiger Kontexte nicht genügend Ansatzpunkte für deren inhaltliche Bewertung – und damit auch für die Durchführung der in Kapitel 4 zu beschreibenden Annotationsaufgabe – bietet. Auf entsprechende Schwierigkeiten geht auch Cramer (2011: 85) im Hinblick auf die von ihr untersuchten Alltagsdefinitionen ein, wenn sie schreibt: „Es ist z. B. völlig unklar, wie praktisch sichergestellt werden kann, dass alle zum Definiendum genannten Merkmale tatsächlich wesentlich für die Definition sind […]“. Relevanter als formale Klassifikationen wissenshaltiger Kontexte scheinen für die in vorliegender Arbeit untersuchten Fragestellungen mithin Ansätze, die begrifflich-theoretische Grundlagen für die Bestimmung des Informations- und Erkenntniswertes wissenshaltiger Kontexte in empirischen Datensammlungen legen. Wird folglich nach einer epistemologischen Klassifikation wissenshaltiger Kontexte, d. h. danach gefragt, welche Informationen und welche Qualität von Informationen durch KRCs vermittelt werden können, ist zunächst auf eine Studie von Bierwisch und Kiefer (1969) zu verweisen, in der die Eigenschaften von Definitionen – gemäß obiger Definition eine Untergruppe von wissenshaltigen Kontexten – aus theoretischer (generativer) Sicht untersucht werden. Bierwisch und Kiefer unterscheiden vier grundsätzliche Arten von Definitionen:

- *Proper definitions*, welche notwendige und ausreichende Attribute eines Begriffs im Sinne der klassischen Terminologielehre und strukturalistischen Semantik benennen: „[… W]e may characterize non-redundant, or proper definitions as those definitional sentences whose definiens provide only specifications pertaining to the core of the term to be defined […]” (ebd.: 73).
- *Complete definitions*, welche alle notwendigen und ausreichenden Attribute eines Begriffs im Sinne der klassischen Terminologielehre und strukturalistischen Semantik benennen (ebd.).
- *Partial definitions,* „[… which] do not completely specify the reading of the term that they can be used to introduce […]” (ebd.: 66).
- *Redundant definitions,* welche auch solche Attribute eines Begriffs benennen, die nicht zu den notwendigen und ausreichenden Attributen im Sinne der klassischen Terminologielehre und strukturalistischen Semantik zählen (ebd.).

Die genannten Arten von Definitionen können zudem auch in Kombination miteinander vorliegen, z. B. als *complete redundant definition* (ebd.: 73). Das besondere Interesse an einer solchen Klassifikation von Definitionen besteht einerseits darin, dass – auf Grundlage der obigen Annahme, derzufolge Definitionen lediglich eine Untergruppe von KRCs konstituieren – sie problemlos auf wissenshaltige Kontexte übertragbar scheint, und andererseits darin, dass sie als Grundlage für eine

von sprachlichen Oberflächenphänomenen unabhängige[21] qualitative Bewertung von KRCs – und damit von späteren Extraktionsergebnissen – herangezogen werden kann. Konkret ist die Terminologie natürlich an möglichst vollständigen und hochrelevanten KRCs interessiert. Werden indes als redundante KRCs solche Kontexte verstanden, die formal und inhaltlich zwar nicht in ein intensionales Definitionsschema passen, jedoch solche Informationen enthalten, die im Sinne der hier noch vorzunehmenden inhaltlich-semantischen Definition (Abschnitt 2.3.3) wissenshaltiger Kontexte ebenfalls als relevant anzusehen sind, da sie wichtige deskriptive Informationen enthalten, können auch diese Aussagen als potentiell nutzbringend angesehen werden. Freilich beseitigt auch die von Bierwisch und Kiefer vorgeschlagene Klassifikation nicht die Schwierigkeit, dass Grenzziehungen zwischen notwendigen und ausreichenden Angaben einerseits sowie redundanter Information andererseits in der Regel ebenso problematisch sind wie die Abgrenzung wissenshaltiger Kontexte von generischen Aussagen (Bierwisch/Kiefer 1969: 74 ff.). Allerdings scheint die feingliedrigere und, wie festgestellt, von sprachlichen Oberflächenphänomenen abstrahierende Typologie dieser beiden Autoren für die empirische Arbeit besser geeignet als das relativ starre und deutlich der klassischen Definitionstheorie verhaftete ISO-Schema.
Von Interesse sind in diesem Zusammenhang auch die Ausführungen Stanaitytės (2005), die sich in ihrer Dissertation mit deutschen Alltagsdefinitionen auseinandersetzt: Zunächst einmal machen ihre Beispiele deutlich, dass die Grenze zwischen denjenigen Äußerungen, die Stanaitytė als Alltagsdefinitionen bezeichnet, und den hier untersuchten KRCs fließend ist[22], wobei einige der von der Autorin gegebenen Beispiele nach den hier noch zu entwickelnden und in späteren Kapiteln auf empirische Daten (Textkorpora) anzuwendenden Selektionskriterien durchaus nicht als valide KRCs angesehen würden – die Übernahme einiger begrifflich-terminologischer (insbesondere der Benennungen „Definiendum“ und „Definiens“) und theoretischer Konstrukte aus der Definitionstheorie sowie von Hypothesen zur sprachlichen Gestalt definitorischer Äußerungen und deren Anwendung auf die Beschreibung von KRCs scheint also insofern gerechtfertigt, als eine scharfe Grenze zwischen Definitionen und KRCs weder theoretisch noch praktisch zu ziehen ist. Des Weiteren führt Stanaitytė (ebd.: 196 f.) aus, dass Vollständigkeit im Sinne Bierwischs und Kiefers für Alltagsdefinitionen keineswegs zwingend sei:

> „In der alltäglichen Kommunikation werden diejenigen Eigenschaften einer Objektklasse aufgeführt, die für den Textproduzenten […] am charakteristischsten erscheinen (und am prototypischsten sind). Diese werden von ihm nach ihrer Wichtigkeit oder ihrem Zutreffen hinsichtlich des Wesens der zu beschreibenden Sache angeordnet; sie sind nicht hierarchisch organisiert. Dadurch wird das ‘Gebot’ der notwendig hinreichenden Bedingungen einer Wesensbeschreibung in Frage gestellt. Die Bewertung und Einordnung dieser Eigenschaften von notwendig bis hinreichend kann in der Alltagskommunikation als subjektiv angesehen werden.“

[21] Walter (2010: 56) diskutiert die Beziehung zwischen linguistischer Struktur und Definitionsqualität knapp unter Bezugnahme auf die von ihm referierten Forschungsarbeiten – darunter auch die Arbeiten Pearsons (1998) und Meyers (2001) –, kann in seiner eigenen Analyse aber keinen direkten Zusammenhang zwischen diesen beiden Größen finden. Auch Stanaitytė (2005: 20) unterstreicht, dass „[…] keine Verbindung zwischen dem syntagmatischen Muster [dem zur Realisierung der Definition verwendeten *knowledge pattern*, vgl. Kapitel 3, Anmerkung der Verfasserin] [… und der] formalen Definitionsart […] festgestellt werden kann“.

[22] Auch Bierwisch und Kiefer (1969: 55 f.) geben Beispiele, die durchaus nicht in das Schema intensionaler oder extensionaler Definitionen passen.

Allerdings, argumentiert die Autorin, bedeute dies keinesfalls, dass redundante Alltagsdefinitionen in ihrer Nützlichkeit gegenüber klassischen (aristotelischen) Definitionen gemindert seien (ebd.: 197):

> „Das Korpusmaterial ließ erkennen, dass auch die kontextuell bedingten Merkmale, die die hinreichenden Bedingungen nicht erfüllen, zur Identifizierung des Bezugsgegenstandes und der Verständigung untereinander beitragen. Aufgrund der Angabe von Eigenschaften hat der Adressat Anknüpfungspunkte, mit deren Hilfe er in seinem mentalen Bedeutungslexikon die gemeinte Bedeutung entweder aktivieren oder neu konstituieren kann."

Diese Gesichtspunkte scheinen auch für KRCs zu gelten, da sie ebenso wie die von Stanaitytė untersuchten Alltagsdefinitionen aus kommunikativen Zusammenhängen – Textkorpora – entnommen werden und da an sie, je nach Kontexttyp, ohnehin – gemäß Definition – lediglich die Anforderung gestellt wurde, Ansatzpunkte für das Verständnis eines unbekannten Begriffs zu liefern, nicht aber, diesen gemäß den Regeln kunstgerechten Definierens vollständig auszudefinieren. Der konkreten Kommunikationssituation geschuldete Besonderheiten, z. B. besondere inhaltliche Hervorhebungen, müssen dabei entsprechend der bisherigen Argumentation keinen Nachteil wissenshaltiger Kontexte darstellen, sondern können bei exhaustiver Betrachtung des interessierenden Gegenstands dazu beitragen, interessante Gesichtspunkte eingehender zu beleuchten. Die Vollständigkeit bzw. Unvollständigkeit im Sinne der Nennung aller notwendigen und hinreichenden Eigenschaften eines Begriffs sind dann nur noch graduelle Qualitätsmerkmale, die in bestimmten Kommunikationssituationen sogar kontraproduktiv wirken können (ebd.: 194 ff.), wobei indes die Vollständigkeit der semantischen Beschreibung durch einen vollständigen oder aber eine Reihe mehrerer unvollständiger, aber komplementärer KRCs erreicht werden kann.

2.3.2.3 Linguistische Klassifikation

Zwar war bereits im vorigen Abschnitt angedeutet worden, dass eine Klassifikation wissenshaltiger Kontexte aufgrund formaler oder formal-linguistischer Kriterien nicht für sinnvoll gehalten wird, da sie an der eigentlich interessierenden Frage – der Qualität der durch automatisch extrahierte Kontexte vermittelten Information – vorbeigeht. Allerdings sind derartige Klassifikationen in der Literatur häufig anzutreffen und sollen daher an dieser Stelle aus Gründen der Vollständigkeit knapp referiert werden.

Eine umfassende linguistische Beschreibung wissenshaltiger Kontexte für das Englische leistet Pearson (1998) anhand mehrerer englischsprachiger Korpora. Zwar unterscheidet sich Pearsons Terminologie von der in dieser Arbeit verwendeten, ihre Forschungsergebnisse sind jedoch auch für das hier verfolgte Vorhaben von Interesse. Pearson unterscheidet u. a. *formal* und *semi-formal defining expositives* (ebd.: 135). Während die Struktur von *formal defining expositives* durch die aristotelische Definitionsformel, derzufolge ein Begriff durch Zuordnung eines Oberbegriffs (*genus proximum*) und Aufzählung relevanter Unterscheidungsmerkmale (*differentia specifica*, vgl. Picht 1997: 34) definiert wird, bestimmt ist, verzichten *semi-formal defining expositives* auf die Nennung des Oberbegriffs (Pearson 1998: 157).

Zur Beschreibung von *simple formal defining expositives* geht Pearson (1998: 136) zunächst von der Definitionsformel *X = Y+distinguishing characteristics* aus, wobei X für den definierten Begriff und Y für das *genus proximum* stehen. Das Gleichheitszeichen symbolisiert ein lexikalisches *hinge*, also ein meist prädikatives sprachliches Element (z. B. *comprise*, *define*, *denote*, *consists*), das die Beziehung

zwischen X und Y expliziert (ebd. 139). Häufig ließen sich diese Verben eindeutig einem Definitions- bzw. Relationstyp zuordnen (ebd. 140, Hervorhebungen im Original)[23]:

> „Sager's assertion that connective verbs are devoid of meaning is slightly misleading and we would argue that connective verbs which we have identified may have meaning in the sense that they may determine the type and content of the defining statement which is being provided. They may point to the provision of intensional information, extensional information, information about synonyms or a combination of all of these. A connective verb such as *to be* will provide intensional information, a connective verb such as *consist* may provide extensional information and a connective verb such as *known as* may provide synonyms. The connective verb may tell us something about the type of distinguishing characteristic which is about to follow. For example, a connective verb such as *be used to* is likely to be followed by a statement of purpose of (sic!) function."

Mit *complex formal defining expositives* benennt Pearson (ebd.: 151-157) definierende und erklärende Textstellen, die sich über mehrere Sätze erstrecken. Diese wurden jedoch aus der Definition wissenshaltiger Kontexte ausgeklammert und sollen demzufolge hier nicht näher betrachtet werden. Eine eingehendere Beschäftigung mit mehrsätzigen Kontexten ist zweifellos von großem wissenschaftlichen Interesse, übersteigt jedoch – auch aus pragmatischen Gründen – den Rahmen dieser Arbeit.

Die von Pearson (ebd.: 157-162) beschriebenen *semi-formal defining expositives* verzichten, wie bereits angemerkt, auf die Nennung eines *genus' proximum*. Darüber hinaus beschreibt Pearson (ebd.: 162-167) sogenannte *dictionary type definitions*, also Definitionen, wie sie in Wörterbüchern und Lexika typisch seien. Diese Definitionen zeichneten sich insbesondere durch typographische Hervorhebungen aus. Wörterbuchdefinitionen können als gut erforscht gelten, auf entsprechende Literatur wurde bereits verwiesen. Auf eine Wiederholung dieser Ausführungen soll hier aus Gründen der Knappheit und Straffheit der Darstellung verzichtet werden.

Tabelle 1 gibt einen Überblick über die von Pearson (1998: 135-167) beschriebenen Strukturtypen einschließlich der von der Autorin angeführten Beispiele aus dem Englischen. Die in Spalte 2 typographisch hervorgehobenen Strukturelemente Definiendum (unterstrichen), Definiens (fett) und *hinge* (kursiv) sind zur leichteren Erkennung auch in den Beispielen entsprechend markiert.

Kontext-Strukturtyp	**Struktur**	**Englisches Beispiel (Pearson 1998)**
a) formal defining expositives		
Simple formal defining expositives	X=**Y**+distinguishing characteristics	Bile *is a* **green liquid** made in your liver and it is stored in your gall bladder.
	Y+distinguishing characteristics =X	**One important natural activity** which can lead to coastal change *is known as* Longshore Drift.
complex formal defining expositives		
This is called ...	defining statement. *This is called ...*	To get pure lines the plants are pollinated with their own pollen. *This is called* self-pollination.

[23] Die gegenteilige Darstellung Stanaitytės (2005: 20) war bereits zitiert worden (Fußnote 21).

heading	heading. *This is ...*	The iris *This is* **a flat ring of muscle** which controls the amount of light that enters the eye.
Zweisatz-Erklärung	Term. *This is a ...*	... a mortgage. *This is a* **long-term loan** either at a fixed or a changing rate of interest.
b) semi-formal defining expositives		
Funktion	Term *is used to/for ...*	Impure salt *is used for* deicing roads.
Meronymie	Term *has/have*	Red cells *have* haemoglobin in them.
c) Dictionary type definitions	Term *Markup* **defining statement**	#GR# Data acknowledgement (AK) #AS# A data acknowldgement message is used to control the window flow control mechanism which has been selected for the data tansfer phase (UDT).

Tabelle 1: Strukturtypen wissenshaltiger Kontexte und weiterer Kontexttypen nach Pearson (1998)

Als alternative Verbindungselemente (*hinges*) können neben Prädikaten auch Bindestriche oder andere typographische Elemente sowie – im Falle von sich über mehrere Sätze erstreckenden Kontexten – kohärenzbildende linguistische Elemente wie Demonstrativpronomina und –adjektive oder sonstige Pronomina fungieren. Von besonderem Interesse ist jedoch Pearsons Gedanke, dass wissenshaltige Kontexte insbesondere durch die Verwendung spezifischer, als Verbindungselement zwischen Definiens und Definiendum fungierender Prädikate charakterisiert sind. Kann diese Sichtweise auch für andere Sprachen und Fachgebiete als die von Pearson untersuchten bestätigt werden, ließen sich wissenshaltige Kontexte anhand entsprechender Prädikate und der von Pearson beschriebenen Strukturen identifizieren.

Studien (für das Katalanische Feliu und Cabré 2002, für das Französische u. a. Malaisé et al. 2005, für das Spanische u. a. Sierra et al. 2008, eine detaillierte Darstellung zu linguistischen Extraktionsmustern für eine Reihe europäischer Sprachen findet sich in Kapitel 3) haben die Existenz ähnlicher wie der in Tabelle 1 präsentierten Strukturen in weiteren europäischen Sprachen untersucht. Angemerkt sei zudem, dass auch in diesen Studien strukturelle Typologien wissenshaltiger Kontexte entwickelt werden, die häufig auch einen „informellen“ Kontexttyp einschließen, aus inhaltlicher Sicht dem bereits Dargestellten aber nichts Wesentliches hinzufügen.

Auch Meyer (2001: 283 ff.) liefert eine Typologie wissenshaltiger Kontexte und unterscheidet ähnlich wie bereits Dubuc und Lauriston (1997) zwischen definitorischen und erklärenden wissenshaltigen Kontexten. Erstere entsprechen im Wesentlichen Pearsons *simple formal defining expositives*. *Dictionary defining KRCs* als Untergruppe definitorischer Kontexte seien in Lehr- und Wörterbüchern anzutreffen, häufig typographisch markiert und generell von hoher Qualität. Dagegen wiesen natürlich vorkommende, also in Texten der natürlichen Alltagskommunikation auffindbare definitorische Kontexte eine Reihe von Unregelmäßigkeiten auf, wie etwa Variationen des Genusterms und in lexikographischen Definitionen unübliche modale Einschränkungen. Die erklärenden KRCs sind dagegen mit Pearsons *semi-formal defining expositives* deckungsgleich. Auch Malaisé et al. (2005) sowie Sierra et al. (2008) präsentieren Kontexttypologien. Allerdings scheinen diese Typologien unterschiedliche Beschreibungsebenen zu vermischen, so dass sie hier nicht berücksichtigt werden.

Hinsichtlich der Wahl der passenden Terminologie erscheinen sowohl Pearson – aus Gründen der Einfachheit und Sprachökonomie – als auch Meyer – wegen terminologischer Überschneidungen mit den Benennungen für formale Kontexttypen – nicht als optimal. Zum Ausdruck des strukturell-

linguistischen Vollständigkeitsaspekts – des Fehlens oder Vorhandenseins eines *genus proximum* – werden hier daher die Termini *Kontext mit/ohne Genusterm* geprägt.
Tabelle 2 präsentiert einen Überblick über die bisher diskutierten Ansätze für die Klassifikation wissenshaltiger Kontexte. Die Benennung der Relationstypen „hierarchisch“ und „assoziativ“ folgt dabei ISO 704: 2009 (ISO 2009b). Eine eingehendere Diskussion der für wissenshaltige Kontexte relevanten Relationstypen folgt in Abschnitt 2.3.3. Wie bereits ausführlich diskutiert, wird die epistemologische Klassifikation als die für das hier verfolgte Forschungsvorhaben sinnvollste angesehen.

Klassifikationskriterium				
nach formalen Gesichtspunkten	definitorischer K.	erklärender K.	assoziativer K.	
nach epistemologischen Gesichtspunkten	proper KRC	complete KRC	redundant KRC	partial KRC
nach semantischer Relation	Kontext mit hierarchischer Relation		Kontext mit assoziativer Relation	
nach Vollständigkeit	Kontext mit Genusterm		Kontext ohne Genusterm	

Tabelle 2: Klassifikation wissenshaltiger Kontexte

2.3.3. Durch wissenshaltige Kontexte verbalisierte semantische Relationen

2.3.3.1. Eigenschaften semantischer Relationen und Bedeutung für die Terminologiearbeit

Bereits in der Einleitung zu dieser Arbeit und auch in den vorstehenden Abschnitten wurden semantische Relationen als wesentliche Elemente wissenshaltiger Kontexte hervorgehoben. Semantische Relationen strukturieren semantisches Wissen über Objekte und Sachverhalte und sind damit neben der Terminologie für so unterschiedliche Disziplinen wie die Informationsextraktion (vgl. Xu 2007) und die künstliche Intelligenz (insbesondere im Zusammenhang mit semantischen Netzen, vgl. Reichenberger 2010) von Bedeutung. Für den Bereich der Lexikographie betonen Kremer und Abel (2010: 380 f.):

> „Semantic relations are relevant for pedagogical lexicography from a variety of perspectives. An extensive presentation of semantic relations within a word entry is important in productive L2-tasks, when a language learner has to choose the semantically correct word out of many possibilities. Furthermore, as the co-occurrence of semantically related items within a text section is probable, in writing tasks the learner should have access to those items. More generally, the presentation of semantic relations offers useful access paths to lexical data in addition to access paths based on merely formal criteria.”

Diese Argumente scheinen auch auf die Terminologiearbeit – zumal in Übersetzungs- und Dolmetschkontexten – zuzutreffen. So betonen Nistrup Madsen et al. (2000: 7) die Bedeutung semantischer Relationen für terminographische Begriffsbeschreibungen: „Giving information about a concept relation and a related concept corresponds to the information on a characteristic of a concept“. Eine umfassende Präsentation semantischer Relationen in Terminologiedatenbanken durch Integration wissenshaltiger Kontexte erleichtert demnach die Auswahl des richtigen Zielterminus‘ und erlaubt den Zugang zu anderen relevanten Begriffen. Zusätzliche Bedeutung gewinnen semantische Relationen, wie bereits einleitend erwähnt, im Zusammenhang mit dem Aufbau terminologischer Wissensbanken. Die Inklusion semantischer Relationen in entsprechende Ressourcen mit Hilfe wissenshaltiger Kontexte könnte demzufolge umfangreiche Möglichkeiten der

konzeptuellen Weiterentwicklung bestehender Terminologiedatenbanken über eine bloße Anreicherung mit zusätzlichen Daten hinaus eröffnen (Auger/Barrière 2008: 1): „Semantic relations are at the core of any representational system, and are keys to enable next generation of information processing systems with semantic and reasoning capabilities [sic!].“ Halskov und Barrière (2008: 21) setzen schließlich wissenshaltige Kontexte explizit zu semantischen Relationen in Beziehung:

> „In the context of computational terminology, especially for the task of generating or updating ontologies and terminological knowledge bases, an important type of gold nuggets are semantic relations expressed explicitly in natural language strings. Such strings have been called knowledge-rich contexts (KRCs).”

Der linguistischen Beschreibung semantischer Relationen wurden in den letzten Jahren zahlreiche Publikationen gewidmet, wobei für Extraktionsexperimente sowohl förderliche als auch nachteilige Eigenschaften semantischer Relationen identifiziert wurden. So nennt Murphy (2003: 26) die folgenden Eigenschaften semantischer Relationen:

- Produktivität,
- Binarität[24],
- Variabilität,
- Prototypikalität[25],
- Semi-Semantizität,
- Unzählbarkeit,
- Vorhersagbarkeit,
- Universalität.

Als für vorliegende Arbeit wichtigste Eigenschaften sind hier die Annahmen der Universalität und Vorhersagbarkeit semantischer Relationen hervorzuheben: Bestimmte, grundlegende Typen semantischer Relationen sind demnach für die Strukturierung semantischen Wissens über Sprachgrenzen hinweg verantwortlich, damit sprachunabhängig und mithin grundsätzlich terminologisch modellierbar: „Relations among words adhere to general patterns, indicating that semantic relations are rule-governed. [...] The same semantic relations are relevant to the description of any language’s lexicon” (ebd.: 26).

Allerdings bergen die von Murphy beschriebenen Eigenschaften semantischer Relationen auch Schwierigkeiten. So beschreibt die Eigenschaft Variabilität den Umstand, dass semantische Relationen sich nur auf einzelne Bedeutungselemente lexikalischer Einheiten, nicht aber auf deren Gesamtbedeutung beziehen können (insbesondere in Fällen von Polysemie). Unter Semi-Semantizität wird verstanden, dass semantische Relationen auch durch sprachliche Faktoren beeinflusst werden und mithin nur unvollständig durch rein semantische Modelle beschrieben werden können.

[24] „Some relations relate only pairs of words, although larger sets of words may be semantically available for the relation” (Murphy 2003: 26).

[25] „Some word sets better exemplify a relation than others, and some word sets [...] seem to have special status as canonical examples of a relation” (Murphy 2003: 26).

2.3.3.2. Entwicklung einer Typologie semantischer Relationen für die computergestützte Terminologiearbeit und die Identifikation wissenshaltiger Kontexte

Insofern als in den bisherigen Ausführungen semantische Relationen als wesentliche Elemente terminographischer Begriffsbeschreibungen und mithin wissenshaltiger Kontexte identifiziert wurden, deren Beschreibungsradius sie gleichsam definieren, ist die Erstellung einer Typologie semantischer Relationen sinnvoll, um Zielrelationen festlegen, Annotationen durchführen und Extraktionsergebnisse evaluieren zu können. Allerdings ist mit Kremer und Abel (2010: 382) davon auszugehen, dass nicht alle theoretisch möglichen semantischen Relationen für die Terminologie generell von Bedeutung sind und unterschiedliche semantische Typen durch unterschiedliche Relationen am besten beschrieben werden: „[...F]or different concept classes different sets of semantic relation types are more salient“. Die Schwierigkeit für die Terminologie besteht mithin darin, jene semantischen Relationen zu definieren, die am besten für die Beschreibung von Begriffen geeignet sind und gleichzeitig sprach- und domänenübergreifend häufig auftreten. Auf die Frage nach diesen Relationen liefern indes weder die terminologische noch die linguistische Fachliteratur eine klare Antwort (Beghtol 2001: 99 f.): „All possible relationships [...] have not been discovered or utilized [...], and no taxonomy of all kinds of relationships has been published.”
Für die Terminologie nennt Wüster (1991[3]: 10 ff.) zwei Typen von Relationen zwischen Begriffen:

- logische Relationen aufgrund der Ähnlichkeit zweier Begriffe,
- ontologische Relationen, die auf zeitlichen, räumlichen oder kausalen Beziehungen beruhen.

Eine schematische Darstellung der Relationstypologie Wüsters bietet Nuopponen (1994: 533):

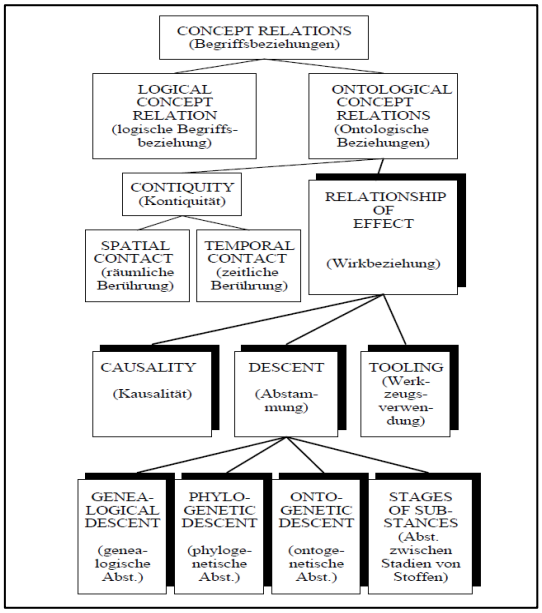


Abbildung 3: Schematische Darstellung der Typologie semantischer Relationen nach Wüster/Nuopponen

ISO 704: 2009 (ISO 2009b) nennt hierarchische sowie assoziative Relationen, wobei allerdings ISO 1087-1: 2000 (ISO 2000: 5) letztere wiederum in temporale, sequentielle und kausale Relationen unterteilt. Nistrup Madsen et al. (2000: 1) untersuchen Typologien semantischer Relationen in der lexikalischen Semantik und der Terminologie und entwickeln eine eigene Typologie, die im Bereich des *information retrieval* zur Anwendung gelangen soll. Die Autorinnen (ebd.: 7 ff.) nennen zunächst assoziative (unspezifische), generische (hierarchische) und partitive Beziehungen sowie die

Relationen *location relation*, *temporal relation*, *purpose relation*, *function relation, role relation*[26] und *property-characteristic relation*, die teilweise auch eine Reihe von Subtypen besitzen (ebd.: 17). Die Autorinnen nennen auch eine *measurement relation*, die allerdings als eine eher spezialisierte, d. h. nicht ausreichend generische Relation hier außer Acht bleiben soll (vgl. auch ebd.: 12). Für die Terminologiearbeit beschreibt zudem ISO 12620: 2009[27] eine lange und durchaus offene Liste von Relationen[28], durch Definitionstypen abgedeckt werden allerdings nur die Relationen Instanz/Beispiel (extensionale Definition), logische Über-/Unterordnung (intensionale Definition) und Meronymie (partitive Definition). Weitere Informationen über möglicherweise ebenfalls relevante Beziehungen können lediglich als zusätzliche Erklärung hinzugefügt werden. Nicht zuletzt beschäftigt sich auch Stanaitytė (2005: 198 f.) mit den Typen semantischer Informationen, die durch Alltagsdefinitionen vermittelt werden. Sie unterscheidet dabei:

- Erläuterungen als eine Umschreibung des Ursprungs,
- Erläuterungen als eine Umschreibung funktionaler Eigenschaften,
- Erläuterungen als eine Umschreibung kausaler Beziehungen,
- Erläuterungen als eine Umschreibung konditionaler Beziehungen (hiermit sind Wenn-dann-Regeln gemeint),
- Erläuterungen als eine Umschreibung verschiedener Operationen,
- Erläuterungen als eine Umschreibung von Form, Gestalt, Erscheinungsbild.,
- temporale und lokale Einordnung.

Des Weiteren beschreibt Stanaitytė (ebd.: 214 ff.) Meronymie- sowie Hyperonymie-/Hyponymiebeziehungen in den von ihr untersuchten Alltagsdefinitionen. Eine Typologie semantischer Relationen legen auch Feliu und Cabré (2002) vor. Die Autorinnen unterscheiden dabei Ähnlichkeit, Inklusion, Sequenzialität, Kausalität, Instrumentalität, Meronymie und Assoziativität mit jeweils einer Reihe von Untertypen.

Die linguistische Fachliteratur zum Thema semantische Relationen beschäftigt sich weitgehend mit den bereits wohlbekannten Relationen Hyponymie/Hyperonymie und Meronymie/Holonymie (vgl. Bean/Green 2001, Green et al. 2002, Murphy 2003) und entwickelt dabei Typologien, die zu feinkörnig für den hier verfolgten Zweck scheinen. Allerdings werden auch in der Computerlinguistik semantische Relationen – insbesondere solche innerhalb von Nominalphrasen – aktiv erforscht. So beschäftigte sich der Semeval07-Task 04 (Girju et al. 2007) mit den folgenden semantischen Relationen:

- Ursache-Wirkung, z. B. „laugh (cause) wrinkles (effect)“,
- Werkzeug-Ausführender, z. B. „laser (instrument) printer (agency)“,
- Produkt-Produzent, z. B. „honey (product) bee (producer)“,
- Begriff-Herkunft, z. B. „message (entity) from outer-space (origin)“,
- Thema-Werkzeug, z. B. „news (theme) conference (tool)“.

---

[26] Diese Relation scheint überwiegend Implikationen für die linguistische Oberflächenstruktur zu haben. Auf semantischer Ebene wird sie als weitgehend äquivalent zu der Relation „Funktion“ aufgefasst und daher hier nicht weiter betrachtet. Als eigene Kategorien scheinen allenfalls die Subkategorien *origin* und *causal relation* von Interesse.

[27] Zitiert nach http://www.isocat.org/interface/index.html, Zugriff am 22.08.2011.

[28] „Concept relations form the basis for concept systems. Types of concept relation can include: generic relation partitive relation sequential relation temporal relation spatial relation associative relation [open list]. [sic!]”

Séaghdha und Copestake (2009) nennen die sechs semantischen Relationen *Be*, *Have*, *In*, *Agent*, *Instrument*, *About*. Am Rande bemerkt sei zudem, dass auch Cimiano (2006: 208 f.) aus ontologischer Sicht unter Bezugnahme auf Aristoteles und das Generative Lexikon von Pustejovsky (1991) auf die sogenannten Qualia-Strukturen *material (constitutive)* (physische Beschaffenheit eines Gegenstands), *agentive* (Ursache für und Prozess der Entstehung eines Objekts), *formal* (die Beschreibung dieser Qualia ist recht unspezifisch, Cimiano (ebd.) verweist ein wenig widersprüchlich sowohl auf die auch von Pustevjovsky (1991: 427) genannte Beschreibung „that which distinguishes the object in a larger domain", unter der dann allerdings im Wesentlichen physische Beschreibungsmerkmale genannt werden, ebenso wie auf Oberbegriffe; tatsächlich scheinen hiermit typologische, also auf Oberbegriffe bezogene Informationen intendiert zu sein) und *final (telic)* (Funktion eines Objekts) verweist und diesen damit gleichsam eine grundlegende ontologische Gültigkeit sowie Generizität zuspricht, die für die hier geführte Diskussion insofern relevant ist, als es hier ebenso wie in Cimianos Arbeit darum geht, allgemeingültige (generische) Strukturen der Wissensrepräsentation zu umreißen (Cimiano 2006: 207, Hervorhebungen durch die Autorin):

> „Qualia structures […] describe a fixed set of relations which *every object* possesses. […] Once we have identified the qualia structure for a given object, we have in fact also identified important ontological properties of this object. "

Tabelle 3 leistet schließlich eine – vereinfachende und vergröbernde - Gruppierung der wesentlichen der in der Literatur beschriebenen Relationen und präsentiert die dieser Arbeit zugrunde liegende Typologie:

Feliu, Cabré	**Girju et al.**	**ISO 1087-1**	**Nistrup Madsen et al.**	**Séaghdha, Copestake**	**Stanaitytė**	**Eigene Typologie**
Inclusion	-	Generisch	Generic	Be	Hyponymie/ Hyperony-mie	Hyponymie/ Hyperony-mie
Meronymy	Werkzeug-Ausführen-der	Partitiv	Partitive	Have	Meronymie/ Holonymie	Meronymie/ Holonymie
Sequentia-lity	-	Temporal	Temporal	-	temporal	Prozess
Sequentia-lity	-	Sequentiell	Location	In	lokal	Position
Causality, Instrument	Ursache-Wirkung, Produkt-Produzent	Kausal	Causal	Agent, Instrument	Kausalität	Kausalität
-	Begriff-Herkunft	-	Origin, material	-	Ursprung	Ursprung
-	Thema-Werkzeug	-	-	-	-	-

Similarity	-	-	Property-Characte-ristic	Be	operativ, Erschei-nungsbild	Referenz
-	-	-		About	-	-
-	-	-	Function, Purpose	-	Funktion	Funktion
Association	-	-	Associative	-	-	-
-	-	-	-	-	Wenn-dann	-

Tabelle 3: Existierende Typologien semantischer Relationen und eigene Zusammenfassung

Es zeigt sich, dass bereits ISO 1087-1: 2000 (ISO 2000) im Zusammenspiel mit Nistrup Madsen et al. (2000) und Stanaitytė (2005) eine gute Grundlage für die Beschreibung semantischer Relationen in wissenshaltigen Kontexten liefert. Computerlinguistische Ansätze, die zur semantischen Klassifizierung von Nominalkomposita dienen sollen, scheinen dagegen weniger relevant. Von Interesse ist jedoch die Relation *Be* nach Séaghdha und Copestake (2009), die sowohl eine generische als auch eine Referenz-Relation (hierunter sollen Aspekte wie etwa Synonymie, aber auch redundante KRCs erfasst werden) darstellen kann – in der hier entwickelten Typologie sollen Referenzrelationen jedoch deutlich von generischen Relationen unterschieden werden: Erstere entsprechen am ehesten dem Typ der Ähnlichkeitsrelation im Sinne von Feliu und Cabré (2002) bzw. der hier zugrunde gelegten Interpretation der von Nistrup Madsen et al. (2000) leider nicht definierten *property-characteristic relation*, sollen aber auch die von Stanaitytė (2005) ins Spiel gebrachten Kontexte operativen und äußerlich bschreibenden Typus‘ umfassen. Die Relation „Ursprung“ umfasst in der hier entwickelten Typologie sowohl Erläuterungen der Herkunft eines Begriffs („kommt von griech. ... und bezeichnet ...“) als auch des Ursprungs gewisser materieller Objekte („wird aus ... hergestellt/gewonnen ...“) und kann zu dem Typus der *origin relation* nach Nistrup Madsen et al. (2000) in Beziehung gesetzt werden, ist allerdings weiter gefasst als diese, da sie auch die von Nistrup Madsen et al. (ebd.) als eigene Beziehung postulierte *material relation* abdeckt. Die Beziehungen *Function* und *Purpose* von Nistrup Madsen et al. (ebd.) werden ebenfalls – durchaus vereinfachend – für den hier zu erfüllenden Zweck als weitgehend deckungsgleich angesehen. Die Relationen *Thema-Werkzeug*, *About*, *Association* und *Wenn-dann* schließlich scheinen für die hier angestrebte generische Typologie semantischer Relationen für terminographische Zwecke nicht relevant.

Insgesamt ist nichtsdestotrotz anzumerken, dass psycholinguistische Studien zur Bedeutung semantischer Relationen für die Wissensstrukturierung – und damit auch für die Wissensrepräsentation in Terminologiedatenbanken – noch am Anfang stehen (vgl. Kremer/Abel 2010). Als hinreichend gesichert gelten lediglich Standardrelationen wie etwa Meronymie/Holonymie (vgl. Pribbenow 2002), so dass zukünftige, psycholinguistisch fundierte Typologien semantischer Relationen durchaus Unterschiede zu der hier entwickelten Darstellung aufweisen können. Bereits Nistrup Madsen et al. (2000: 12) unterstreichen zudem: „The relations not considered central for terminology work in this overview may very well turn out essential when analysing the terminology of some specialised domain“ (z. B. in den Fachgebieten Medizin oder Recht). Dieser offensichtliche Umstand kann auch im Angesicht der hier entwickelten generischen, d. h. robusten Typologie semantischer Relationen für die Terminologiearbeit nicht genug betont werden. Andererseits ist auch anzumerken, dass einige der für die Beschreibung semantischer Beziehungen in der Allgemeinsprache diskutierten Relationen wie etwa die Oppositionsbeziehung (vgl. Löbner

2003: 121 ff.) oder Wortfelder (ebd.: 128 ff., vgl. auch Herbst/Klotz 2003: 72) eher aus linguistischer Sicht von Interesse sind, für die terminologische Begriffsbeschreibung jedoch nur eine untergeordnete Rolle zu spielen scheinen und daher hier vernachlässigt werden können. Die zuletzt genannten Beziehungstypen wurden bereits in Abschnitt 2.2 als lexikalische Relationen von dem hier interessierenden Untersuchungsgegenstand der wissenshaltigen Kontexte abgegrenzt.

2.4 Zusammenfassung

In diesem Kapitel wurde der Begriff des wissenshaltigen Kontexts als des zentralen Untersuchungsgegenstands dieser Arbeit theoriegeleitet geschärft. Zunächst wurden wissenshaltige Kontexte von anderen Kontexttypen theoretisch abgegrenzt und in Bezug auf ihre spezifische terminographische Funktion sowie die sich daraus ergebenden obligatorischen Klassenmerkmale diskutiert. Hierauf aufbauend konnte alsdann die von Meyer (2001) stammende Definition des *knowledge-rich context* erweitert und präzisiert werden. Am Rande wurde dabei auch die Beziehung wissenshaltiger Kontexte zu Definitionen thematisiert: Einerseits haben vorstehende Ausführungen gezeigt, dass nicht alles, was in der zitierten Literatur als Definition bezeichnet wird – das Spektrum reicht von juristischen Definitionen (Walter 2010) bis hin zu sogenannten Alltagsdefinitionen (Stanaitytė 2005, Cramer 2011) – durch den in der Terminologie verwendeten Begriff der Definition abgedeckt wird, so dass die Unterscheidung zwischen Definitionen und wissenshaltigen Kontexten durchaus schwerfällt. Andererseits wurden Definitionen, insbesondere aristotelische Definitionen – bewusst vereinfachend – als Sonderformen bzw. Spezialfälle wissenshaltiger Kontexte aufgefasst, die mithin strukturell weniger strengen Vorgaben unterliegen als Definitionen selbst. Im Zusammenhang mit der Diskussion unterschiedlicher Klassifikationsschemata für KRCs wurde zudem herausgearbeitet, dass inhaltlich-epistemologische Klassifikationsansätze von größerer Bedeutung für vorliegende Arbeit sind, wobei darauf hingewiesen wurde, dass auch die Orientierung an den semantischen Merkmalen klassischer Definitionen – die Nennung der notwendigen und hinreichenden Begriffsmerkmale – nicht nur an in der Praxis schwer überprüfbare Anforderungen geknüpft ist, sondern für die Beschreibung von Begriffen im hier intendierten Sinn ein zu enges Korsett darstellt, da angenommen wird, dass die kommunikative Wirklichkeit deutlich vielfältiger und variantenreicher ist als die terminologische Definitionstheorie (die hier, dies soll nicht verheimlicht werden, allerdings nur in Ansätzen referiert wurde). Die zitierten Forschungsarbeiten scheinen diese Schlussfolgerung im Großen und Ganzen zu stützen. Um dennoch die Grundlagen für die Annotation wissenshaltiger Kontexte in deutschen und russischen Textkorpora zu schaffen, wurde abschließend eine Typologie der durch wissenshaltige Kontexte verbalisierten semantischen Relationen entwickelt. Im folgenden Kapitel wird der aktuelle Forschungsstand zu wissenshaltigen Kontexten dargestellt.

3. Wissenshaltige Kontexte: Stand der Forschung

In der Einleitung zu dieser Arbeit waren die folgenden Forschungsfragen aufgeworfen worden:

- Wie wird fachspezifisches Wissen in Texten dargestellt, d.h. welche linguistischen Merkmale weisen wissenshaltige Kontexte auf?
- Welche Arten von fachspezifischem Wissen sind relevant für die übersetzungsorientierte Terminologiearbeit, d. h. welche Arten fachspezifischen Wissens sind in Terminologiedatenbanken darzustellen, um den Nutzern solcher Ressourcen einen effizienten und strukturierten, übersetzungsorientierten Wissenserwerb zu ermöglichen?
- Wie erkennen Übersetzer fachspezifisches Wissen in Texten, d. h. gibt es überhaupt einen Konsens in Bezug auf die Relevanz fachspezifischer Aussagen in Texten für die übersetzungsorientierte Terminologiearbeit?
- Wie kann relevantes semantisches Wissen in natürlichsprachigen Texten mit Hilfe von Computerprogrammen erkannt und automatisch aus ihnen extrahiert werden?
- Inwieweit sind automatisch aus Korpora extrahierte wissenshaltige Kontexte für die Terminologie überhaupt nutzbar?

Der aktuelle Forschungsstand zu diesen einzelnen Themenkomplexen soll nun in möglichst knapper Form referiert werden. Auf Frage Nummer 2 wird dabei jedoch nicht weiter eingegangen, da die vorstehenden Ausführungen gezeigt haben dürften, dass sowohl in der terminologischen Praxis als auch in der wissenschaftlichen Fachdiskussion noch kein hinlänglich konsolidierter und erst recht kein kodifizierter Wissensstand zu diesem Punkt existiert. Aus diesem Grund war in Kapitel 2 versucht worden, eine (wenngleich vereinfachende, so doch) handhabbare und theoretisch fundierte Definition wissenshaltiger Kontexte einschließlich einer Typologie der durch sie vermittelten Informationen zu entwickeln. Diese Frage wird somit hier im Zusammenhang mit dem aus der wissenschaftlichen Fachliteratur entnehmbaren Forschungsstand nicht weiter diskutiert. Auf Frage Nummer 5 können schließlich erst im weiteren Verlauf dieser Untersuchung sowie an deren Ende Antworten gegeben werden. Vor der Darstellung des Forschungsstandes soll jedoch ein allgemeiner Überblick über die mit dem Thema insgesamt befassten Disziplinen gegeben werden.

3.1 Allgemeiner Forschungsstand: disziplinäre Eingrenzung

Roelcke (2013: VII) beschäftigt sich mit der auch für vorliegende Arbeit relevanten Frage, „[…] wie terminologische Systeme per definitionem in fachlichen Texten eingeführt werden“. In der Tat ist hiermit eine Fragestellung formuliert, die sehr nah an dem mit dieser Dissertation verfolgten Forschungsinteresse liegt, allerdings in ihrer Breite noch über dieses hinausgeht. Obwohl Roelckes Arbeit in methodologischer Hinsicht auf eine Untersuchung von Frequenzen auf Einzelwortebene beschränkt bleibt, ist seine Darstellung des relevanten Forschungsstandes in der Fachsprachenlinguistik, der Terminologie, der Fachlexikographie und Terminographie etc. doch von Interesse und soll daher hier kurz referiert werden.

Für den Bereich der Fachsprachenlinguistik konstatiert Roelcke (ebd. 6), dass „[…] Strategien der Vertextung von terminologischen Systemen […] im Hintergrund“ bleiben. Für die germanistisch ausgerichtete Fachsprachenlinguistik sei eine Konzentration auf textlinguistische Aspekte

festzustellen. Auch die Fachtextlinguistik selbst habe bislang nur wenig Interesse an der textuellen Einbettung terminologischer Systeme gezeigt.
Für die Terminologie selbst stellt Roelcke (ebd.: 7 f.) fest, „[…] dass das Interesse der Terminologielehre wiederum vornehmlich den Bezeichnungen und Bedeutungen einzelner Termini sowie terminologischen Systemen gilt, wobei […] Normierung und Dokumentation im Vordergrund stehen“: „Strukturen oder Strategien der Konstituierung von Terminologie in Texten bleiben indessen weitgehend unberührt, obwohl sie einen wesentlichen Beitrag zum Verständnis von Termini leisten.“ Dies gelte auch für die Tätigkeit von Standardisierungsorganisationen wie ISO und DIN, allerdings ignoriere ein solcher Ansatz den aktuellen Stand der Fachsprachenlinguistik:

> „Der Grund für die Orientierung an einzelnen Termini und Terminologien mag im Falle der Terminologielehre und der Terminologienormung in deren Sprachkonzeption liegen, die als systemlinguistisches Inventarmodell charakterisiert werden kann […]. Hiernach setzen fachlexikalische bzw. terminologische Regelungen ausschließlich am lexikalischen Inventar und am lexikalischen System an und lassen dabei pragmatische oder gar kognitive Gesichtspunkte weitgehend außer Acht. Ein solches Modell beherrscht die Terminologielehre und –normung bis heute, obwohl die Fachsprachenlinguistik bereits seit Längerem mit einem pragmalinguistischen Kontext- oder einem kognitionslinguistischen Funktionsmodell arbeitet.“

Diese prononciert vorgetragene Kritik bedarf freilich einiger ergänzender Bemerkungen. So ist es zwar korrekt, dass im Bereich der Terminologie empirische, korpusbasierte Ansätze bislang nur eine untergeordnete Rolle spielen. Allerdings sollten die bisher zitierten genuin terminologischen Arbeiten sowie die Sonderausgabe „Pattern-based Approaches to Semantic Relation Extraction“ der Zeitschrift „Terminology“ (14 (1), 2008), die Roelcke (2013: 6) auch nennt, gezeigt haben, dass eine gewisse fachinterne Dynamik durchaus existiert. Kritikwürdig bleibt jedoch zum einen auch angesichts dieser Fortschritte, dass korpusbasierte Techniken in der praktischen Terminologiearbeit nach wie vor kaum Anwendung finden, was wohl auch an der nach wie vor ungenügenden theoretischen und methodischen Elaboriertheit ersterer liegen mag. Zu bemängeln ist zum anderen das fast vollständige Fehlen theoretischer Rückkopplungen, die sich aus solchen Untersuchungen ergeben könnten. Eine Ursache hierfür könnte wiederum darin liegen, dass Brückenschläge zu Nachbardisziplinen, die sich mit ähnlichen Fragestellungen beschäftigen, etwa der Computerlinguistik oder aber der Germanistik und allgemeinen Linguistik – auch aus diesen Bereichen enthält vorliegende Arbeit zahlreiche Literaturnachweise –, bislang kaum zustande gekommen sind und entsprechende Forschungsergebnisse nur schwach rezipiert werden.
Relevante Beiträge zur Untersuchung der Konstituierung terminologischer Systeme verortet Roelcke (ebd.) in der Fachlexikographie und Terminographie (Stichwort Terminologisierung durch lexikographische Nachschlagewerke), der Definitionslehre (Stichwort Klassifikation von Definitionen) und der lexikalischen Semantik (Stichworte *folk taxonomies,* Frame-Semantik). Insofern, als diese Aspekte für vorliegende, korpusbasierte Untersuchung relevant sind, wurde auf sie bereits hingewiesen. Darüber hinaus sollen sie an dieser Stelle aber nicht thematisiert werden. Insgesamt ist festzustellen, dass Begriffserklärungen und Definitionen in Texten in verschiedenen Disziplinen untersucht werden, dabei allerdings kaum interdisziplinäre Kommunikation stattzufinden scheint, so dass die einzelnen, sich in methodologischer und auch theoretischer Hinsicht durchaus unterscheidenden, jedoch immer noch genügend Ansatzpunkte für einen erfolgreichen Wissenstransfer bietenden Ansätze in ihren jeweiligen Einzeldisziplinen weitgehend isoliert bleiben. Zu Recht verweist Roelcke zudem auf die kognitive Dimension der Thematik, die – soweit bekannt –

im Rahmen korpuslinguistischer bzw. empirisch-linguistischer Ansätze bislang nicht beleuchtet wurde.

3.2 Linguistische Merkmale wissenshaltiger Kontexte

Linguistische Eigenschaften von wissenshaltigen Kontexten und Definitionen über die in Abschnitt 2.3.2.3 hervorgehobenen Klassifikationsmerkmale hinaus sind in der Literatur insbesondere im Hinblick auf spezifische linguistische Muster bzw. *knowledge patterns* beschrieben worden, anhand derer relevante Textbestandteile automatisch zu identifizieren seien. Des Weiteren wurden informationsstrukturelle Aspekte, semantische Typen von Definitionsprädikaten sowie lexikalische, phraseologische und syntaktische Merkmale wissenshaltiger Kontexte und Definitionen untersucht, wobei allerdings für das Russische nach Kenntnis der Autorin keinerlei Daten vorliegen und die Übertragbarkeit der von Stanaitytė (2005), Walter (2010) und Cramer (2011) präsentierten Ergebnisse für das Deutsche aufgrund der starken inhaltlichen Spezialisierung der Untersuchung Walters sowie der Konzentration Cramers und Stanaitytės auf allgemeinsprachliche Definitionen auf die hier untersuchten wissenshaltigen Kontexte zu terminologischen Begriffen fraglich ist. Die Untersuchungen Oberholzers (2002) und Storrer/Wellinghoffs (2006) schließlich scheinen für generelle Rückschlüsse als zu wenig umfangreich.

3.2.1 Definitionen und wissenshaltige Kontexte als sublanguage

Obwohl im letzten Kapitel Wörterbuchdefinitionen aus dem engeren Gegenstandsbereich der vorliegenden Arbeit ausgeschlossen wurden, ist die Arbeit Barnbrooks (2002) aufgrund ihrer recht detaillierten Darstellung dennoch von Interesse. Barnbrook (ebd.: 72 ff.) untersucht englischsprachige Definitionen in einem einsprachigen Wörterbuch, wobei er diese Definitionen als eigenes Sprachregister (*sublanguage*) auffasst und folglich mittels einer eigens entwickelten Definitionsgrammatik untersucht. Unter Bezugnahme auf die entsprechende Fachliteratur postuliert Barnbrook (ebd.: 75 f.) die folgenden Merkmale von *sublanguages*:

- eingegrenzter Gegenstandsbereich,
- lexikalische, syntaktische und semantische Einschränkungen,
- spezifische grammatische Regeln,
- hohe Frequenz bestimmter Strukturen,
- spezielle Textstruktur,
- Verwendung von Sonderzeichen.

Freilich können diese Merkmale auf Fachsprachen generell bezogen werden. Beispielsweise nennt Barnbrook (ebd.: 79 f.) eine stark ausgeprägte *lexikalische* Uniformität – gemessen am Verhältnis der Wortzahl in einem Korpus zur Anzahl der lexikalischen Typen (Lemmas)[29], der Anzahl der *hapax legomena* sowie der *uni-gram*-Entropie (ebd.: 80) – definitorischer Texte als Indikator für das Vorliegen eines eigenen Sprachregisters. Als *syntaktischer constraint* wird die Aussagesatzform genannt (ebd.: 80). Eine weitere Besonderheit – und diese Besonderheit kann tatsächlich nicht ohne Weiteres auf Fachsprachen generell übertragen werden, wohl ist aber ihre Gültigkeit für Definitionen und KRCs außerhalb von Wörterbüchern zu überprüfen – sei das weitgehende Fehlen von Kohäsionsmitteln, Referenzoperationen und Ellipsen (ebd.: 81):

[29] Vergleiche auch die Diskussion zur *Type-Token-Ratio* in Abschnitt 4.1.4.1.

> „Each unit of definition text […] is a single sentence […]. These sentences are almost completely independent of each other within the dictionary text, so that virtually none of the normal cohesive devices found in all connected forms of text occur […]. Because of this, almost all references to entities already introduced into the text operate entirely within the individual sentence. Ellipsis, the use of pronouns to replace repeated elements and the other normal features of cohesion are only found on a very limited scale and are almost always contained within the single definition sentence.“

Auf genuin syntaktische, d. h. auf konkrete Konstruktionen bezogene Einschränkungen verweist Barnbrook (ebd.: 83 f.) dagegen nur summarisch:

> „[… I]t is evident […] that the range of possible sentence structures, and hence the syntactic range of the language used within the sentences, is significantly restricted as compared with the language at large.“

Auf *semantischer* Ebene postuliert Barnbrook (ebd.: 84) eine Tendenz zur Verwendung generischer Lexik („[…] tendency towards the most general use of defining words […]“). Als *textuelle* Merkmale nennt Barnbrook schließlich – ähnlich wie bereits Pearson (1998, vgl. Abschnitt 2.3.2.3) – die Position von Definiens und Definiendum sowie typographische Markierungen, die für vorliegende Arbeit weniger relevant sind, da sie außerhalb von Wörterbüchern zweifellos in anderer Art und Weise verwendet werden. Barnbrooks spezifische *grammatische Strukturen* rekurrieren auf die Extraktionsgrammatik, welche der Autor im Zuge seiner Arbeit entwickelt. Unter *hochfrequenten Strukturen* versteht der Autor insbesondere bestimmte Konstruktionstypen lexikographischer Definitionen, die er in einer eigenen Taxonomie beschreibt. Die Beispiele 1) und 2) (ebd.: 135 f., Hervorhebungen im Original) entsprechen einigen dieser Typen:

1) **Forgot** is the past tense of **forget**.
2) You can also say you **admire** something when you look with pleasure at it.

Die genannten *Sonderzeichen* – insbesondere für lexikographische Angaben wie etwa Wortart, Genus etc. – und wörterbuchspezifischen *Textstrukturen* schließlich scheinen für das hier verfolgte Interesse bestenfalls von untergeordneter Bedeutung zu sein.

3.2.2 Weitere linguistische Merkmale

Walter (2010: 40) beklagt in seiner Dissertation zu juristischen Definitionen das Fehlen systematischer linguistischer Untersuchungen insbesondere auf statistischer Ebene:

> „In der linguistischen Literatur fehlen systematische statistische Untersuchungen zu Definitionen bisher völlig. […] Qualitativ zeigt sich […] jedoch deutlich, dass sowohl die Vorgehensweise als auch die Formulierungsmuster […] oft deutlich von den in der Wissenschaftstheorie sowie der Terminologiewissenschaft und Lexikographie diskutierten Schemata ‚kunstgerechten‘ Definierens abweichen.“

In seiner eigenen Untersuchung (ebd.: 218, 269 ff.) beschreibt Walter allerdings Eigenschaften deutscher juristischer Definitionen, anhand derer Extraktionsergebnisse mit Hilfe eines Verfahrens des maschinellen Lernens gerankt werden können. Für vorliegende Untersuchung von Interesse sind dabei:

- Lexikalische Merkmale:
    - Boostwörter wie „Merkmal“ oder „Begriff“, die unabhängig vom Definitionsprädikat auf den Definitionscharakter der Aussage verweisen,
    - Stopwörter wie beispielsweise deiktische Adverbien, Definita, Demonstrativpronomen oder Adjektive mit demonstrativer Funktion, die nur selten in Definitionen erscheinen,
    - Negationen, die nur in Ausnahmefällen in Definitionen verwendet werden,
    - Modifizierende Satzadverbien in kontextualisierten Definitionen,
- Referentielle Merkmale, z. B. Oberbegriff oder Synonym des Definiendums im Definiens,
- Strukturelle Merkmale, z. B. Position des Definiendums vor dem Definiens, Definitionsprädikat im Hauptsatz,
- Dokumentbezogene Merkmale, z. B. relative Position der Definition im untersuchten Dokument, weitere Treffer im Umfeld der Definition,
- Sonstige Merkmale, z. B. Satzlänge, Relevanz-Score des Definiendums[30].

Walter (ebd.: 64) stellt zudem fest, dass das Definiendum „[…] durchgängig einem Argument des Definitionsprädikats […]“ entspricht und nennt das Passiv als „[…] Zusatzindikator für das Vorliegen einer Definition (ebd.: 169). Cramer (2011: 185 ff.) diskutiert zusätzlich zu den bislang genannten Merkmalen eine Reihe weiterer Charakteristika von Definitionen, die auch für die Beschreibung wissenshaltiger Kontexte von Interesse sein können. Zunächst bestreitet sie aufgrund ihrer Analysen die in der Literatur häufig anzutreffende Behauptung, dass definitorische Aussagen überwiegend an exponierten Textpositionen, beispielsweise am Beginn oder Ende eines Textes zu finden seien, verweist aber darauf, dass definitorische Aussagen zur Bildung von *clustern* oder Reihen neigen, also in Texten konzentriert auftreten. Des Weiteren führt Cramer (ebd.: 188 f.) Beispiele für lexikalische und phraseologische Merkmale definitorischer Aussagen an. Diese lassen sich den folgenden Gruppen zuordnen:

- generische Nomina, z. B. „Definition“, „Begriff“, „Bedeutung“,
- Fachgebiets- und Herkunftsangaben, z. B. „in der Mathematik“, „aus dem Lateinischen“,
- Generalisierungssignale, z. B. „in der Regel“, „typischerweise“, „gängig“,
- Adverbien, z. B. „definitionsgemäß“, „fachsprachlich“, „wörtlich“,
- Ausdrücke der Unsicherheit und Floskeln des Abwägens und Relativierens, die indes ein kontroverses Merkmal darstellen, da beispielsweise Pearson (1998: 141) Kandidaten mit entsprechenden Formulierungen aus ihrer Untersuchung ausschließt – Cramer dagegen beschäftigt sich mit alltagssprachlichen Definitionen, für die entsprechende Floskeln möglicherweise ein geradezu typisches Merkmal darstellen. Der Status linguistischer Relativierungssignale kann damit nicht abschließend geklärt werden.

Personalpronomen und überhaupt Mittel des linguistischen Ausdrucks von Subjektivität und persönlicher Involviertheit führt Cramer (2011: 151, 188) dagegen als Negativmerkmal bzw. Stopwort an, bemerkt allerdings in Bezug auf lexikalische Merkmale generell (ebd.: 189): „Leider sind viele lexikalische Merkmale bzgl. des Merkmalstyps recht häufig, aber in ihrer konkreten Merkmalsausprägung selten“. Für das Englische referiert Pearson (1998) sowohl Ansätze zur Beschreibung von Definitionen aus älteren Studien, insbesondere zur Verwendung bestimmter Definitionsprädikate, zur grammatischen Form des Definiendums, zur Verwendung des unbestimmten Artikels sowie zu unterschiedlichen Definitionstypen (ebd.: 89 ff.), als auch zur

[30] Walter nennt mit dem TF-IDF-Score ein in der Computerlinguistik häufig verwendetes Maß für die Relevanzbewertung von Termini (vgl. Manning/Schütze 1999: 541 ff.).

Beschreibung der Definitionen, die sie im Rahmen ihrer eigenen Untersuchungen in Textkorpora identifiziert. Einige der von Pearson identifizierten strukturellen Klassifikationsmerkmale von Definitionen in Textkorpora wurden bereits in Kapitel 2 beschrieben. Weitere linguistische Merkmale englischer *formal defining expositives* sind zudem nach Pearson:

- Verwendung des unbestimmten Artikels oder keines Artikels (aber möglicherweise eines anderen Determinans') bei satzinitialer Position des Definiendums (ebd.: 138),
- bei nicht-satzinitialer Position des Definiendums Einschränkung der erlaubten Determinatoren auf den bestimmten oder unbestimmten Artikel (ebd.),
- Vorliegen der Definitionsbeziehung „Definiendum = Definiens" im Hauptsatz (ebd.: 141),
- Voranstellung von Gültigkeitsangaben in Form eines Nebensatzes (z. B. „In biology, ...", „In science, ...", ebd.),
- Anfügen der *differentia* an das Definiens mit Hilfe einer Präposition, eines Relativpronomens oder eines past participle (ebd.: 144).

Fahmi und Bouma (2006: 65) diskutieren syntaktische Eigenschaften niederländischer Definitionen, gehen dabei aber von der vereinfachenden Annahme aus, dass „[… p]otential definition sentences are sentences containing a form of the verb *zijn* (to be) with a subject and nominal predicative phrase as sisters", also weitestgehend Definitionen aristotelischen Typus' entsprechen. Die Autoren nutzen des Weiteren eine Liste von Stopwörtern, um falsche Positive auszuschließen:

> „[… I]f the subject is headed by (the Dutch equivalents of) *cause*, *consequence*, *problem*, *result*, *feature*, *possibility*, *symptom*, *sign*, etc., or contains the determiner *geen* (no), the sentence will not be included in the list of potential definitions."

Für *machine-learning*-Experimente verwenden die Autoren auch eine Reihe von *features*, von denen das Gros allerdings nicht im hier intendierten Sinn interpretierbar ist (z. B. sogenannte *bag-of-words*-Ansätze). Allerdings verweisen Fahmi und Bouma (ebd.: 67) auch auf syntaktische Regularitäten: „[… A] definition sentence with a copular pattern tends to put its subject in the beginning." Zudem diskutieren die Autoren die bereits von Pearson (1998) beschriebenen Regularitäten bei der Verwendung von Artikeln („[… T]he majority of subjects in definition sentences have no determiner […]"), und verweisen darauf, dass definitorische Sätze mehr Eigennamen enthalten (ebd.).

Westerhout (2009) schließlich beschreibt – basierend auf den Konnektoren, die Definiendum und Definiens miteinander verbinden – eine Reihe unterschiedlicher Definitionstypen, die allerdings der bisherigen Darstellung nichts Neues hinzuzufügen vermögen. Auch Westerhout diskutiert alsdann eine Reihe linguistischer Merkmale – Artikelgebrauch, Verteilung von Eigennamen –, die ebenfalls für *machine-learning*-Experimente genutzt werden sollen und im Einklang mit den bisherigen Ausführungen stehen. Interessanterweise untersucht Westerhout (ebd.: 64) auch typographische Hervorhebungen und kommt zu dem Ergebnis, definitorische Aussagen wiesen anteilig mehr solcher Hervorhebungen auf. Westerhout (ebd.: 65) untersucht Positionsmerkmale definitorischer Aussagen, insbesondere die Position definitorischer Aussagen in Textabsätzen (drei der vier untersuchten Definitionstypen tendieren dazu, eher am Beginn eines Absatzes aufzutreten), steht mit ihren Ausführungen allerdings in einem tendentiellen Widerspruch zu Cramer (2011). In ihrer im Zusammenhang mit hybriden Ansätzen weiter unten diskutierten Dissertation (Westerhout 2010) entwickelt die Autorin ihren Ansatz weiter.

Marshman (2008) untersucht in ihrer Studie insbesondere *expressions of uncertainty*, also relativierende Ausdrücke, mit deren Hilfe die Gültigkeit bzw. Relativität der durch einen KRC vermittelten Information markiert werden können. Die Autorin untersucht insbesondere die

Phänomene Quantifizierung (Einschränkung der Gültigkeit auf eine Teilmenge der Begriffsextension), lexikalische Unsicherheitsmarker (*hedges*), Modalverben und Negation. Marshman findet quantifizierende Angaben lediglich in sechs (für einen französischen Datensatz) bzw. sieben Prozent (für einen englischen Datensatz) der von ihr untersuchten Daten. *Hedges* treten in den von der Autorin untersuchten Daten deutlich häufiger auf, und zwar in 15 % (für das Französische) bzw. 24 % (für das Englische) der untersuchten KRCs, im Schnitt etwas unterhalb dieser Werte lagen auch die beobachteten Häufigkeiten von Modalverben und Negation. Marshman diskutiert die genannten Phänomene insbesondere als Merkmale, welche die Verwendbarkeit von KRCs oder vielmehr KRC-Kandidaten einschränken und geht darin mit Pearson (1998, s. o.) konform. Andererseits beschreibt etwa Cramer (2011) – für das Deutsche – relativierende Elemente als geradezu typisch für Alltagsdefinitionen. Gemäß der für diese Arbeit in Kapitel 2 entwickelten Definition wissenshaltiger Kontexte kann sich diesem Standpunkt aber nicht angeschlossen werden, vielmehr wird der Sichtweise Pearsons (1998) und Marshmans (2008) gefolgt. Bei der Erstellung der empirischen Datenbasis wird mithin darauf zu achten sein, dass Kandidaten mit allzu starken Relativierungen nicht berücksichtigt werden.

3.2.3 Knowledge patterns

Musterbasierte Verfahren der Extraktion wissenshaltiger Kontexte sowie des *ontology learning* und der Extraktion semantischer Relationen gehen auf eine Pionierstudie von Hearst (1992) zurück, in der nicht-prädikative lexikalisch-syntaktische Muster zur Extraktion englischer Nominalphrasen für die automatisierte Bildung einer Taxonomie verwendet wurden. Beispiele hierfür sind die Ziffern 3)-6). Hearsts Beispiele sind dabei ähnlich vereinfachten regulären Ausdrücken, wie sie in der Programmierung verwendet werden, zu lesen: der Ausdruck *{NP,}** bedeutet daher „0 oder eine bis mehrere Nominalphrasen, gefolgt von einem Komma", *{(or|and)}* bezeichnet ein disjunktes ODER, also das Vorkommen entweder von *and* oder *or*. Die englischen Beispiele stammen ebenfalls von Hearst.

3) such NP as {NP,}*{(or|and)} NP
   ... works by such authors as Herrick, Goldsmith, and Shakespeare.
   Ordnet die Autorennamen dem Genusterm „Autor" unter.
4) NP {, NP}*{,} or orther NP
   Bruises, wounds, broken bones or other injuries ...
   Ordnet die Bezeichnungen für die Verletzungen unter den entsprechenden Genusterm unter.
5) NP {,} including {NP,}*{or|and} NP
   All common-law countries, including Canada and England ...
   Ordnet die Ländernamen unter die erste Nominalphrase unter.
6) NP {,} especially NP {,}*{or|and} NP
   ... most European countries, especially France, England, and Spain.
   Ordnet die Ländernamen unter die erste Nominalphrase unter.

Im Zusammenhang mit wissenshaltigen Kontexten hat sich für solche lexikalisch-syntaktischen Definitions- bzw. Erklärungssignale der Terminus *knowledge pattern* durchgesetzt, den Marshman (2007: 4) auf Meyer (u. a. 2001) zurückführt. Condamines (2002: 145) verwendet dagegen den Terminus *conceptual relation pattern*, den sie als „[… a] discursive structure used as an indication of

the possible transition from the discourse to a model, allowing the more or less direct construction of a model in the form of a semantic relation depending on its relation with the context“ definiert. Stanaitytė (2005: 16 ff.) spricht von „syntagmatischen Mustern“, die sie als „[…] formelhafte Syntagmen, die stets in Verbindung mit Alltagsdefinitionen stehen und daher als ihre Indikatoren verstanden […]“ werden können, definiert. Allerdings abstrahiert sie kaum von sprachlichen Oberflächenstrukturen, womit notwendigerweise eine Loslösung der fraglichen Realisierungsmuster von der Ebene des Syntagmas verbunden wäre. Da eine solche Lösung einerseits konzeptuell und auch theoretisch sinnvoller erscheint und andererseits sprachliche Muster (wie von Condamines 2002 diskutiert) nicht direkt auf semantische Relationen abgebildet werden können, wird in vorliegender Untersuchung für die Bezeichnung solcher Realisierungsmuster ebenfalls der englische Terminus *knowledge pattern* (im Gegensatz zu *conceptual relation pattern*) oder aber die deutsche Benennung „Extraktionsmuster“ (alternativ ggf. „Suchmuster“) gewählt.

Pearson (1998: 148 ff., 158 und 182 ff.) nennt wie auch andere Autoren in anderssprachigen Studien (z. B. Sierra et al. 2008, vgl. auch Gerzymisch-Arbogast 1996) für das Englische insbesondere syntaktische Prädikate, mit deren Hilfe relevante Definitionsbestandteile üblicherweise miteinander verknüpft werden. Ähnliche Musterkataloge mit ebenfalls meist prädikativen Elementen finden sich in den Arbeiten von Condamines und Rebeyrolle (1998 und 2001 – Englisch), Feliu und Cabré (2002 – Katalanisch), Malaisé et al. (2005 – Französisch), Sierra et al. (2008 – Spanisch), Oberholzer (2002 – Deutsch), Stanaitytė (2005 – Deutsch), Storrer und Wellinghoff (2006 – Deutsch), Walter (2010 – Deutsch) sowie Cramer (2011 – Deutsch). Tabelle 4 gibt einen Überblick über wesentliche Elemente dieser Musterkataloge. Die lexikalisch-syntaktischen Muster sind dabei in der von den jeweiligen Autoren gewählten Form wiedergegeben, wobei deutlich fachgebietsspezifische oder aber nicht eindeutig zuordenbare bzw. nicht zu den gewählten semantischen Relationen passende Muster nicht berücksichtigt werden. Einzig Cramers, Stanaitytės und Oberholzers Muster werden in vereinfachter Form dargestellt, da Cramer in ihren Musterbeschreibungen auch strukturelle Aspekte berücksichtigt und Oberholzer stark auf syntaktische Aspekte abhebt, wohingegen Stanaitytė unterschiedliche grammatische Realisierungsformen ein und desselben Musters beschreibt, die hier (noch nicht) interessieren: Diese Aspekte sollen von der lexikalischen Komponente getrennt behandelt werden.

Condamines und Rebeyrolle (2001: 138) nennen ebenfalls eine Reihe sequentieller lexikalischer Muster für das Englische wie „is composed of“ oder „has responsibility for“. Hierbei handelt es sich allerdings um Übersetzungen von Musterbausteinen, die in einem französischen Korpus gefunden wurden. Ähnlich wie Oberholzer (2002) beschreiben zudem auch Condamines und Rebeyrolle syntaktische Strukturen wissenshaltiger Kontexte. Da der Status der übersetzten Muster in der Arbeit von Condamines und Rebeyrolle indes nicht unmittelbar einsichtig ist, werden ihre Muster nicht in die Tabelle aufgenommen. Für solche Muster der Tabelle, die nicht intuitiv verständlich sind, enthält eine Parenthese einen Hinweis auf die intendierte Lesart. Semantisch ähnliche Muster sind in der Tabelle zusammengefasst, wobei bewusst von syntaktischen Unterschieden – etwa der Unterscheidung Pearsons (1998) zwischen Hauptsätzen und Appositionen – abstrahiert wurde, da diese für die hier interessierenden Sprachen Russisch und Deutsch ohnehin gesondert zu untersuchen sind. Spalte 9 enthält gleichzeitig die in Abschnitt 2.3.3.2 (Tabelle 3) erarbeiteten Zielrelationen, denen die einzelnen Muster zugeordnet wurden, auch wenn diese Zuordnung im Einklang mit der bisherigen Darstellung (Polysemie von Extraktionsmustern) bestenfalls als tendenziell zu bezeichnen ist.

Tabelle 4 zeigt erhebliche Überschneidungen, aber auch deutliche Unterschiede zwischen den einzelsprachlichen Untersuchungen auf: In Abhängigkeit davon, welche semantischen Relationen für die Untersuchung ausgewählt werden, scheinen die einzelsprachlichen Listen ineinander

konvertierbar und aus der Gesamtheit Musterlisten scheinen sich durch Übersetzung Listen für das Russische generieren zu lassen. Gleichzeitig enthalten die Musterkataloge z. T. auch solche Muster, die mit der in Kapitel 2 entwickelten Definition wissenshaltiger Kontexte und der dort erarbeiteten Typologie semantischer Relationen nicht kompatibel sind, so dass diese Elemente nicht in die Tabelle aufgenommen wurden. Dies und der Umstand, dass die Relation Prozess mit nur einem Muster, die Relationen Position, Ursprung und Kausalität mit sehr wenigen Mustern vertreten sind, zeigt, dass die Beschreibung sprachlicher Muster stark von der Definition des Untersuchungsgegenstands und dem gewählten Korpus abhängig ist. Eine Bewertung der Vollständigkeit eines Musterkatalogs ist demnach in Abhängigkeit von der zugrunde gelegten Definition wissenshaltiger Kontexte durchzuführen. Umfangreiche empirische Untersuchungen wie etwa die Walters (2010) zeigen zudem, dass in der Praxis eine deutlich größere Bandbreite von *knowledge patterns* auftritt als in kleineren oder eher theoriegeleiteten Studien (ebd.: 60, 79). Unterschiedliche Autoren – etwa Stanaitytė (2005: 17) sowie Storrer und Wellinghoff (2006) und Sierra et al. (2008: 92) – verweisen darüber hinaus auf die schwankende Zuverlässigkeit von Extraktionsmustern. Tabelle 4 zeigt zudem, dass auch nicht-prädikative Elemente, etwa bestimmte Nomina, Adjektive oder Elemente, die Nebensätze oder Parenthesen einleiten, in der Literatur als Musterbestandteile beschrieben werden, zusätzlich zu diesen nennen etwa Malaisé et al. (2005: 34) und Cramer (2011: 188) auch Elemente der Zeichensetzung, die auf das Vorliegen definitorischer Aussagen hinweisen. Przepiórkowski et al. (2007) beschreiben Extraktionsgrammatiken für das Bulgarische, das Tschechische und das Polnische. Die Autoren nennen unterschiedliche Strukturtypen von Definitionen, beispielsweise für das Bulgarische *NP – NP* (eine auch im Russischen mögliche Variante, in der die Kopula durch einen Bindestrich ersetzt wird). Bulgarische *knowledge patterns* sind:

- представлява (stellt dar),
- показва (zeigt),
- означава (bedeutet),
- описва (beschreibt),
- се използва (wird verwendet (für)),
- позволява (erlaubt),
- дава възможност да (ermöglicht),
- се нарича (wird genannt),
- подобрява (verbessert),
- осигурява (stellt sicher, garantiert, gewährleistet),
- служи за (dient (dazu)),
- се разбира (wird verstanden (als)),
- обозначава (bezeichnet),
- съдържа (enthält),
- определя (definiert, bestimmt),
- включва (einschließen),
- се дефинира като (wird definiert (als)),
- се основава на (gründet auf, basiert auf).

Die Autoren nennen auch tschechische Extraktionsmuster auf Grundlage ähnlicher Prädikate:

- definuje (definiert),
- znamená (bedeutet),
- vymezuje (grenzt ab, definiert),
- představuje (stellt dar).

Pearson	Feliu / Cabré	Malaisé et al.	Sierra et al.	Oberholzer	Stanaitytė	Storrer / Wellinghoff	Walter	Cramer
								Ursprung
-	-	-	-	-	-	-	-	kommen aus (Sprache)
-	-	-	-	-	-	-	-	sich bilden, wenn
-	-	-	-	-	-	-	-	entstehen
								Hyperonymie
include(s), contain(s)	incloure	-	incluir, comprender	-	-	-	umfassen	-
-	-	-	contar	-	-	-	gehören, zählen (zu)	-
-	-	-	-	-	-	-	-	fallen unter
-	-	une sorte de	-	-	-	-	-	Art/Typ sein, Beispiel
								Referenz
-	-	-	-	-	-	begreifen als	gelten, ansehen, betrachten, anerkennen (als)	-
be defined as	-	définir	definir como	-	definieren	definieren als	-	-
is/are called	-	appeler, baptiser, dénommer, nommer	llamar, lombrar	nennen	nennen, sprechen von, heißen	nennen, heißen, sprechen von	sprechen (von), umschreiben	nennen, sprechen von, heißen

This is	representar-se amb	-	ser, representar, constituir	sein	nichts anderes sein als	sein, spezifizieren als, vorstellen als	sein, darstellen	-
is/are charac-terized by	caracteritzar-se per	-	caracterizar como	-	-	charakterisie-ren als	gekennzeich-net sein	charakterisiert sein/werden (durch), gekenn-zeichnet sein, sich auszeichnen durch, erkennt man daran (, dass)
is/are described as	-	-	describir como	-	-	beschreiben	-	-
known as	-	-	conocer como	-	-	bekannt als	-	-

the term	-	appellation, acception, concept, dénomi-nation, désignation, expression, mot, nom, notion, terme, vocable + appliquer, donner, employer, prendre, porter, recevoir, référer, renvoyer, réserver, utiliser	-	bezeichnen	bezeichnen, Bezeichnung/ Wort/Begriff sein (für), Wort/Begriff steht für, Wort/Begriff bedeutet, hinter dem Begriff verbirgt sich, Definition sein	bezeichnen als, bedeuten, benennen, Terminus einführen	bedeuten, besagen	bezeichnen, Begriff/ Fachwort bezeichnet, Begriff/ Fachwort verwendet man für, meinen
-	asemblar-se a	signifier, vouloir dire, indiquer, équivaloir à, employer pour, marquer, expliquer, préciser, dénoter, designer	referir a, significa, equivaler, ser_igual, ser_similar	-	-	-	-	gleichen, ähneln, entsprechen

-	-	entendre par	comprender como, entender como, concebir	-	verstehen unter	verstehen unter	verstehen (unter)	verstehen unter
-	-	-	-	-	-	-	vorliegen	vorliegen, wenn
-	-	-	-	-	-	-	-	auffassen als
-	il s'agit de	-	-	-	-	-	sich handeln um	sich handeln um
-	-	-	-	-	-	-	-	übersetzen als
i.e., e.g.	c'est-á-dire, en d'autres termes, soit, à savoir, enfin, par exemple, autrement dit, même chose	-	-	-	also, zum Beispiel, zu gut deutsch, zu deutsch, besser bekannt als, zu neudeutsch, neudeutsch, sozusagen, nämlich, sprich	-	also, und damit, sogenannt, das heißt	also, sprich, kurz, d.h., z. B., etwa, zum Beispiel, u.a., unter anderem
-	-	-	-	-	-	-	-	bei … hat man es zu tun mit
-	-	-	-	-	-	-	-	bei[31]
-	-	-	-	-	sogenannt, so genannt	-	sogenannt	-
						-		**Meronymie**
has	-	-	-	-	-	-	-	-
involve(s)	-	-	contener, integrar	enthalten, beinhalten, umfassen	-	-	-	-

[31] Cramer (2011) verweist mehrfach auf diese Präposition in satzinitialer Position mit einer Vielzahl nachfolgender Strukturen.

-	constar de	-	consistir, constar	bestehen aus, sich zusammen-setzen aus	-	bestehen aus	-	sich zusam-mensetzen aus, bestehen aus
-	-	-	constituir	-	-	-	-	(Haupt)-Bestandteil sein
								Funktion
is/are used for/to	servir com	-	emplear, usar, utilizar, emplearse	-	-	verwenden als, gebrauchen	-	einsetzen, verwenden, dienen
-	fer que	-	-	-	-	-	-	-
-	servir com a	-	permitir, servir	-	-	-	-	-
								Prozess
-	augmentar	-	-	-	-	-	-	-
								Kausalität
-	dependre de	-	-	-	-	-	-	-
-	ser la causa (principal) de	-	-	-	-	-	-	-
-	provocar	-	-	-	-	-	-	-
-	contribuir a	-	-	-	-	-	-	-
-	-	-	-	-	-	-	folgen (aus)	-
-	produir	-	-	-	-	-	-	-
								Position
-	tenir lloc a nivel de	-	-	-	-	-	-	-
-	situar sobre	-	-	-	-	-	-	-

Tabelle 4: Aus der Literatur entnommene Elemente für knowledge patterns

Allerdings sind lexikalisch-syntaktische Muster an sich keineswegs unproblematisch: Auf ihre variable Zuverlässigkeit war bereits hingewiesen worden. Bereits Meyer (2001: 295 ff.) verweist zudem auf die Unvorhersehbarkeit (und die daraus resultierende Unmöglichkeit, mit Hilfe von Suchmustern alle in einem Text vorhandenen KRCs zu identifizieren), Polysemie (bezogen auf semantische Relationen) und Fachgebietsabhängigkeit (vgl. hierzu auch die Diskussion in Abschnitt 2.3.3.2) von *knowledge patterns*. Bei Verwendung expressiver Ausdrucksweisen oder zahlreicher Anaphora stößen lexikalisch-syntaktische Extraktionsmuster ohnehin an ihre Grenzen. Auch Condamines (2002) leistet eine ausführliche Diskussion von Besonderheiten, die bei der Verwendung lexikalisch-syntaktischer Extraktionsmuster zu beachten sind (z. B. Korpusgebundenheit bestimmter Muster, Polysemie etc.).

3.2.4 Zusammenfassung

Die obigen Ausführungen zu linguistischen Merkmalen wissenshaltiger Kontexte vermitteln ein ambivalentes Bild. Hier ist zunächst die Charakterisierung von KRCs als eigenes Sprachregister zu diskutieren, die theoretisch auf einer recht schwachen Grundlage steht, da eine Abgrenzung von anderen, übergeordneten Kategorisierungsebenen (z. B. der Ebene der sprachlichen Varietäten insgesamt, denen KRCs zweifellos unterzuordnen sind) aufgrund der von Barnbrook genannten Kriterien fraglich scheint und zudem die Übertragbarkeit von Barnbrooks Thesen auf definitorische Aussagen in natürlichen Verwendungszusammenhängen, d. h. außerhalb von Wörterbüchern auch dann noch in Zweifel steht, wenn bedacht wird, dass Barnbrooks Daten zumindest solche Definitionen enthalten, die in Form ganzer Sätze formuliert sind[32]. Andererseits argumentiert auch Cramer (2011: 149 f.) – mit leicht abweichender Terminologie – im Sinne Barnbrooks: „Interpretiert man Definitionen als eine Art Textsorte, so scheint es […] in der Tat ein gemeinsames Textsortenwissen zu geben.“ Inwieweit die hier noch durchzuführende empirische Untersuchung Indizien für eine solche Abgrenzung erbringt, wird in den folgenden Kapiteln dieser Arbeit zu diskutieren sein.

Des Weiteren zeigt der Literaturüberblick, dass bisherige Forschungsarbeiten weitgehend auf die Untersuchung potentieller Extraktionsmuster konzentriert waren, wobei stark unterschiedliche Niveaus sowohl der technisch-methodologischen als auch der inhaltlich-theoretischen Elaboration der Ansätze festzustellen sind. Häufig werden Untersuchungen in – bewusst oder unbewusst – reduktionistischer Manier auf bestimmte Definitionstypen eingegrenzt. Häufig (z. B. bei Fahmi/Bouma 2006, Storrer/Wellinghoff 2006, Fišer et al. 2010) scheint es sich hierbei um den aristotelischen Definitionstypus zu handeln, dessen Prävalenz in wissenschaftlichen Studien allerdings durch empirische Daten nicht zu rechtfertigen ist. So stellt etwa Cramer (2011: 161) fest, dass „[d]er starke Fokus der Literatur auf den Definitionstyp der aristotelischen Definition […] als zu einseitig“ erscheint. Walter (2010: 57) kommt im Ergebnis seiner Analyse von Begriffsdefinitionen in deutschen Urteilsbegründungen zu einem ganz ähnlichen Schluss: „Das aristotelische Definitionsschema spielt […] nicht die prominente Rolle, die ihm in der Terminologiewissenschaft generell zugewiesen wird.“ Stanaitytė (2005) schließlich präsentiert in ihrer Dissertation zahlreiche Beispiele für Alltagsdefinitionen in journalistischen Artikeln, die den Anforderungen an aristotelische Definitionen ebenfalls nicht entsprechen. Allerdings verweist sie (ebd.: 190) auch darauf, dass Autoren von Alltagsdefinitionen eine Präferenz für den aristotelischen Definitionstypus zeigen:

[32] Wörterbuchdefinitionen wurden bereits aus dem Gegenstandsbereich vorliegender Arbeit ausgeschlossen. In Kapitel 4 wird diese Entscheidung weiter differenziert.

> „Tatsächlich lässt sich anhand von Korpusdaten nachweisen, dass aristotelische Definitionen in alltagsprachlichen Texten häufig anzutreffen sind. Die Fülle an Belegen legt die Vermutung nahe, dass diese Definition aus der wissenschaftlichen Praxis in die alltäglichen Definitionsverfahren von Schreibenden übernommen worden ist. Die vielen Textbeispiele, die der Definitionsform des Genus Proximum entsprechen, lassen Rückschlüsse darüber zu, dass der Textproduzent davon ausgeht, auf diese Weise beim Textrezipienten am schnellsten das Identifizieren des Erläuterungsobjektes zu ermöglichen."

Auch Cramer (2011: 158 ff.) stellt fest, dass – obwohl im Alltag viele unterschiedliche Definitionstypen zum Einsatz kommen – aristotelische Definitionen als besonders hochwertig angesehen werden. Für das hier verfolgte Forschungsvorhaben lässt sich daraus der Schluss ableiten, dass diese Form – sie wurde in Kapitel 2 als Sonderform wissenshaltiger Kontexte beschrieben – unbedingt zu berücksichtigen ist, andere Formen wissenshaltiger Kontexte aber ebenfalls für die terminographische Arbeit zu erschließen sind.

Insgesamt ist Walter (2010) hinsichtlich des von ihm beklagten Fehlens systematischer Untersuchungen der linguistischen Eigenschaften von Definitionen und wissenshaltigen Kontexten zuzustimmen. Auch Untersuchungen, die nicht auf nur einen bestimmten Definitions- oder Kontexttyp beschränkt sind, liefern häufig nur punktuelle und einander zum Teil widersprechende Beobachtungen (z. B. im Hinblick auf das Auftreten einschränkender und relativierender Formulierungen). Viele dieser Beobachtungen beziehen sich zudem lediglich auf die Ebene des Einzelwortes. Befunde beispielweise zu Regularitäten beim Artikelgebrauch oder dem Auftreten von Eigennamen scheinen tatsächlich eher auf *ad-hoc*-Beobachtungen zu beruhen als auf systematischen Studien und können insbesondere aufgrund des Fehlens sprachübergreifender Studien kaum in den Rang bestätigter Regularitäten erhoben werden.

Als relativ gut untersucht können dagegen – mit den oben genannten Einschränkungen – sprachliche Muster für die Extraktion wissenshaltiger Kontexte gelten. Aufgrund der vorauszusetzenden empirischen Variation bei der Formulierung russischer und deutscher wissenshaltiger Kontexte ist zwar bei einer rein musterbasierten Extraktion mit eher mittelmäßigen Performanzwerten zu rechnen, obiger Musterkatalog bietet aber bereits eine gute Grundlage für weitergehende Untersuchungen.

3.3 Annotation wissenshaltiger Kontexte in Textkorpora

Für eine systematische korpuslinguistische Untersuchung linguistischer Eigenschaften wissenshaltiger Kontexte ebenso wie für Extraktionsexperimente stellen im Hinblick auf ihren KRC-Status annotierte Korpusdaten in Form eines sogenannten Goldstandards eine wesentliche Grundlage dar. Aus diesem Grund sollen Annotationsmethodologien und Ergebnisse früherer Annotationsstudien hier kurz referiert werden. Anzumerken ist, dass Studien mit spezifischer Übersetzungsorientierung nicht vorzuliegen scheinen, Antworten auf die Frage, in welchem Maße zwischen Übersetzern in Bezug auf KRC-Kandidaten mit konsensuellen Entscheidungen gerechnet werden kann, also an dieser Stelle nicht gegeben werden können.

3.3.1 Durchführbarkeit der Annotation und Einflussfaktoren

Bierwisch und Kiefer (1969: 56) argumentieren, dass […] definitional sentences cannot be identified on the basis of their syntactic form alone". Die Autoren stellen mithin die konsensuelle Annotierbarkeit wissenshaltiger Kontexte anhand syntaktischer (und möglicherweise allgemein

linguistischer) Kriterien grundsätzlich infrage. Sie (ebd.: 74) führen zudem aus, dass der Unterschied zwischen einer generischen, nicht-definitorischen und einer definitorischen Aussage „[…] on the overall structure of the dictionary which imposes the partition of the reading of the entry in question into core and periphery“ beruhe. Die Unterscheidung generisch-definitorisch wird in dieser Sichtweise folglich durch das lexikalische System derjenigen Sprache oder Sprachvarietät determiniert, zu welcher der Definitions- bzw. KRC-Kandidat gehört. Konsequent zu Ende gedacht bedeutet dies: Was eine Definition oder ein valider wissenshaltiger Kontext ist, ist wiederum Definitionssache bzw. – für den Kontext der Terminologiearbeit – abhängig von dem konkreten Erkenntnisinteresse, das die Terminologiearbeit lenkt, und von der konkreten Sprache oder Sprachvarietät, für die diese durchgeführt wird. Eine zuverlässige Annotation wissenshaltiger Kontexte in Textkorpora kann also nur dann gewährleistet werden, wenn die Annotatoren über eine detaillierte, gemeinsame und explizite Repräsentation dieses Erkenntnisinteresses verfügen – eine nur mit Mühe zu garantierende und noch schwerer zu überprüfende Bedingung.
Cramer (2011) führt in ihrer Dissertation umfangreiche Annotationsexperimente durch und liefert im Ergebnis dieser Arbeit eine detaillierte Darstellung derjenigen Faktoren, die den Annotationsprozess jenseits eines terminologischen Interesses ebenfalls beeinflussen können (ebd.: 181 f.):

> „Wie die praktische Annotationsarbeit zeigt, trifft jeder Annotator seine Entscheidung nicht allein auf Grund der beobachteten Formulierung oder der genannten Informationen. Intuitiv bewertet er zusätzlich eine Reihe weiterer Faktoren, die je nach Gewichtung des jeweiligen Annotators mehr oder weniger stark in die Annotationsentscheidung einbezogen werden. Hierzu zählen z. B., ob die seiner Meinung nach zentralen Merkmale des Definiendums genannt werden und welchen Informationsbedarf der Produzent bei den anvisierten Rezipienten annimmt; daneben spielt sein Textverständnis, sein Vorwissen im Bezug auf das Textthema und die Einstellung gegenüber dem Text sowie die dem Produzenten zugestandene Glaubwürdigkeit eine entscheidende Rolle.“

Abbildung 4 (ebd.: 182) fasst diese Faktoren in Form eines Schaubildes zusammen.

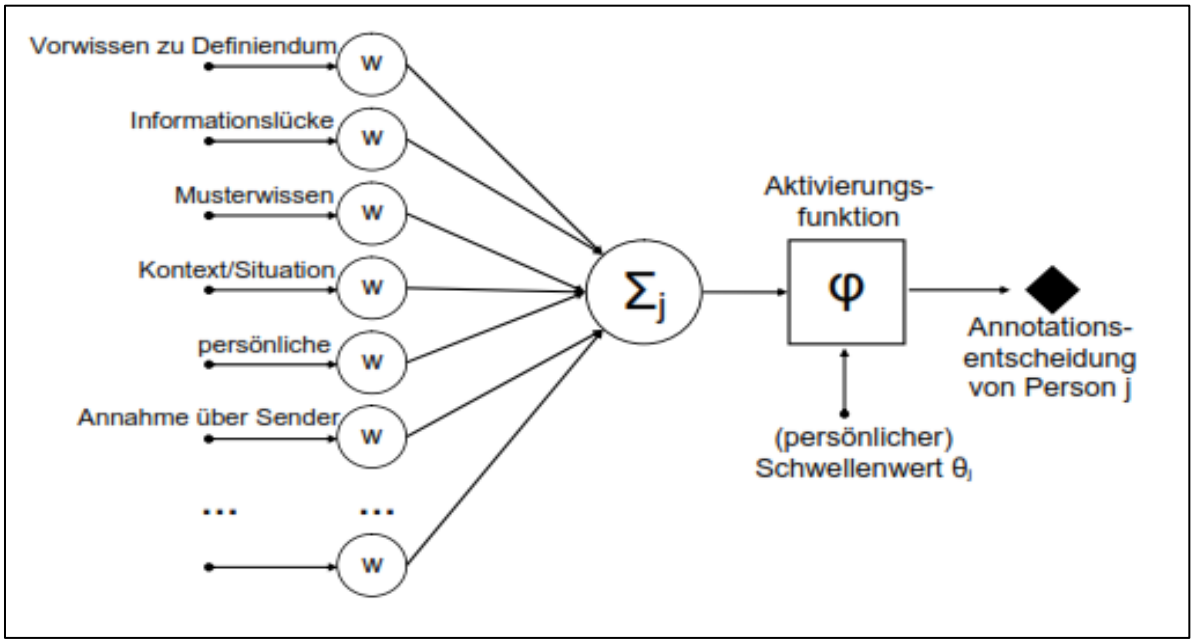


Abbildung 4: Einflussgrößen bei der Annotation definitorischer Aussagen nach Cramer (2011: 182)

Übereinstimmend mit dieser Darstellung kommt auch Walter (2010: 51) zu dem Schluss, dass „[…] der Übergang zwischen definitorischen und nicht-definitorischen Aussagen […] fließend [ist] und […]

auf einem Bündel graduierbarer, teils der semantischen, teils der eher pragmatischen Ebene zuzuordnender Merkmale" basiert. Für vorliegende Arbeit bedeuten diese Ergebnisse, dass eine Annotation wissenshaltiger Kontexte in Textkorpora zwar möglich ist, die zugrundeliegenden Annahmen und die auf die Annotationsergebnisse einwirkenden Einflussfaktoren jedoch transparent gemacht werden müssen.

3.3.2 Annotationsmethodologien

3.3.2.1 Annotationsverfahren

In einer Reihe von Arbeiten zur Beschreibung und Extraktion von definitorischen Aussagen – Definitionen und wissenshaltigen Kontexten – wird offenbar auf die Nutzung eines nach formalen Kriterien erstellten Goldstandards verzichtet, d. h. die Autoren präsentieren und diskutieren diejenigen Beispiele, die sie selbst den jeweils verwendeten Kategorien zuordnen (z. B. Condamines/Rebeyrolle 1998, Pearson 1998, Meyer 2001, Stanaitytė 2005, Fišer et al. 2010). Malaisé et al. (2005: 39) verweisen lediglich darauf, dass „[t]o measure recall[33], we have humanly analysed and tagged the test corpus and a random sample of the evaluation corpus […]“. Del Gaudio und Branco (2007) führen gleichfalls eine manuelle Annotation von Definitionen in ihrem Untersuchungskorpus durch, ohne jedoch – abgesehen davon, dass sie die Häufigkeit jedes vorab definierten linguistischen Definitionstypus‘ in ihrem Korpus angeben – die Zusammenhänge und Hintergründe des Annotationsverfahrens zu erläutern. Przepiórkowski et al. (2007) verweisen auf die manuelle Annotation von 150 Definitionen für jede der von ihnen untersuchten Sprachen, gehen auf die Details des Annotationsprozesses jedoch ebenfalls nicht ein. Marshman (2008) schließlich wendet zwar Termextraktion und Konkordanzverfahren an, um eine Liste von KRC-Kandidaten zu generieren, die Auswahl dieser Kandidaten selbst scheint jedoch nicht das Ergebnis eines formalisierten Annotationsprozesses darzustellen. Sierra et al. (2008: 91) merken zunächst an:

> „The evaluation of a system to extract definitional knowledge is not an easy task […].A ‘gold standard‘ is quite difficult to establish, while the definition of a term could be more or less relevant depending on the specialised level of the evaluated texts as well as the evaluator criteria.“

Der Evaluation ihres Extraktionssystems legen die Autoren dabei scheinbar ebenfalls eine mehr oder weniger subjektive Gütebewertung der extrahierten KRC-Kandidaten zugrunde und verzichten aufgrund des Fehlens von Informationen über *alle* in den untersuchten Korpora auffindbaren KRCs auf die Angabe eines Werts für die Abdeckung der Extraktion (*recall*). Westerhout (2009: 62) schließlich verweist lediglich auf „[…] 603 manually annotated patterns […]“ im Zusammenhang mit der Beschreibung der Datengrundlage ihrer Studie.

Muresan und Klavans (2002) verwenden im Gegensatz dazu eine einfache, inhaltlich nicht näher spezifizierte Annotationsmethode als Grundlage für die Evaluation ihrer Extraktionsexperimente: Vier Personen markieren in jeweils neun medizinischen Artikeln Definitionen und deren Definienda. Die Annotatoren sind dabei keine Medizin-Experten und waren auch nicht an der Entwicklung des Extraktionssystems beteiligt. In den Goldstandard werden alsdann all diejenigen Definitionskandidaten aufgenommen, die von mindestens drei der vier Annotatoren als valide Definitionen markiert wurden. Auf diese Weise entstand ein Goldstandard mit einem Umfang von allerdings nur 53 Definitionen.

[33] Zu relevanten Evaluationsmaßen siehe Abschnitt 3.3.2.2.

Storrer und Wellinghoff (2006) beschränken ihre Untersuchung, wie bereits angemerkt, auf aristotelische Definitionen und entwickeln ein Schema für die XML-basierte Annotation definitorischer Aussagen in Textkorpora, demzufolge sowohl das Definiendum als auch das Definiens und das jeweils realisierte *knowledge pattern* annotiert werden. Als Grundlage der Annotation dient ein Korpus von 20 Dokumenten (ca. 100.000 Wörter) aus den Bereichen Texttechnologie und Hypertext. Die Annotationsregeln werden in einer entsprechenden Richtlinie festgehalten. Die Autorinnen führen dazu aus (ebd.: 2374).

> „We tested our guideline on interoperability with two students. This study revealed that, in some cases, one needs to be familiar with the domains of text technology and hypertext research in order to decide whether a term occurs as a definiendum or is just used in a more general way. Since the annotated definitions are used as the gold standard in our evaluation study, we needed to obtain a reliable basis. Thus, the final decision whether a text segment was annotated as a definition or not was made by a researcher familiar to the domains of text technology and hypertext research."

In Übereinstimmung mit der Argumentation im vorigen Abschnitt verweisen mithin auch Storrer und Wellinghoff (ebd.) darauf, dass für Annotationsentscheidungen letztlich Fachwissen erforderlich ist. Allerdings lässt auch diese Studie offen, welche Annotationsregeln konkret in den Annotationsrichtlinien festgelegt wurden und welchen fachlichen Hintergrund die annotierenden Studenten hatten. Anders als in der Studie von Muresan und Klavans (2002) wird die Entscheidung über die Aufnahme eines Definitionskandidaten jedoch nicht an ein Mehrheitsvotum geknüpft, sondern an Expertenwissen.

Walter (2010) präsentiert eine im Vergleich zu vorgenannten Arbeiten detaillierte Beschreibung seiner Annotationsarbeit. Als Textgrundlage wurden dabei 60 Urteilsbegründungen deutscher Gerichtsentscheidungen unterschiedlicher Rechtsgebiete (275.901 Wörter, ebd.: 73) ausgewählt. Definitorische Passagen wurden dabei mit den Konfidenzbewertungen „sicher", „unsicher" oder „zweifelhaft" markiert (ebd., ebd.: 39), darüber hinaus wurde bei der Annotation zwischen Kern- und elaborierenden Informationen unterschieden und die argumentative Funktion der Sätze bestimmt. Bemerkenswert an diesem Vorgehen ist auch, dass nicht nur Einzelsätze, sondern auch ganze Satzkomplexe annotiert wurden (ebd.: 40). Annotationskriterien wurden in einer recht umfangreichen Annotationsrichtlinie festgehalten[34], in der auch Beispiele und Problemfälle diskutiert werden. Für die Unterscheidung zwischen Definitionen und generischen Sätzen formuliert diese Richtlinie die Entscheidungskriterien Verallgemeinerbarkeit, Essentialität und Relevanz der Information. Das Definiendum müsse zudem als Tatbestandsmerkmal einer Norm fungieren oder für die Auslegung eines solchen von Bedeutung sein. Für die Annotation des Goldstandards wird die Annotation sowohl von Walter selbst als auch von einem Fachgebietsexperten (einem Juristen nach dem 1. Staatsexamen) durchgeführt (ebd.: 73). Die Zusammenführung der Annotationen beider zu einem Goldstandard erfolgt danach konsensuell. Auf diese Weise entsteht ein insgesamt 274 Definitionskomplexe enthaltender Goldstandard mit insgesamt 507 Sätzen (ebd.: 74). Die Qualität der Annotation bewertet Walter – gemessen an dem sogenannten κ-Wert als einem quantitativen Maß zur Überprüfung der Übereinstimmung zwischen Annotatoren[35] – als in einem mittleren, akzeptablen Bereich liegend (ebd.: 75 ff.). Des Weiteren gibt Walter (ebd.: 77 ff.) auch den Grad der Übereinstimmung des Goldstandards mit einem kommerziellen Register von Definitionen in juristischen Texten an. Im Ergebnis resümiert der Autor (ebd.: 79) zwar, dass die Schwierigkeit der Annotationsaufgabe aufgrund der bereits erwähnten großen linguistischen Variationsbreite bei der

[34] http://www.coli.uni-saarland.de/projects/corte/walter06richtlinien.pdf, Zugriff am 21.03.2014.

[35] Die Evaluation manueller Annotationen wird im folgenden Abschnitt diskutiert.

Formulierung definitorischer Aussagen sowie der Komplexität der inhaltlich-funktionalen Verflechtung dieser Aussagen mit dem Gesamtkontext des sie umgebenden Texts beträchtlich sei, aber dennoch Kriterien formuliert und linguistische Signale identifiziert werden könnten, auf deren Grundlage die Unterscheidung definitorisch-generisch mit insgesamt befriedigender intersubjektiver Generalisierbarkeit getroffen werden könne.

> „Bei begriffsbestimmenden Aussagen in Gerichtsurteilen handelt es sich nur zu einem Teil um vollwertige Definitionen im Sinne der Äquivalenz- und Essentialitätsbedingungen […]. Auch das in der Terminologiewissenschaft als prototypisch für Definitionen betrachtete Formulierungsmuster *Ein A ist ein B, das C* wird bei weitem nicht in allen dieser Aussagen verwendet. Begriffsbestimmungen erfüllen unterschiedliche Aufgaben […]. In diesem erweiterten Sinne definitorische Information kann nicht aufgrund strikter Kriterien charakterisiert werden, sondern durch ein Bündel graduierbarer Merkmale. Ob eine Äußerung demnach als definitorisch anzusehen ist, kann zwar unter Umständen nur im Kontext und unter Einbeziehung von Welt- und Fachwissen beurteilt werden. Dennoch haben wir in einer Annotationsstudie zwischen einem juristisch ausgebildeten Annotator und einem Nicht-Juristen eine insgesamt befriedigende Übereinstimmung bei der Identifikation definitorischer Information im weiteren Sinne festgestellt. […] Es ist also anzunehmen, dass stilistische Eigenheiten einzelner Urteilstexte die Identifikation definitorischer Information erleichtern oder erschweren können.“

Die innerhalb der hier betrachteten Literatur umfangreichste Annotationsstudie wurde von Cramer (2011) vorgelegt. Konkret kann Cramer auf die Annotationsarbeiten von Storrer und Wellinghoff (z. B. 2006) aufbauen und vier Annotationsexperimente durchführen sowie Informationen sowohl zu Sichtweisen der Produzenten als auch zu Sichtweisen der Rezipienten von Definitionen erheben (Cramer: 136):

> „[… D]er Fokus der ersten beiden Experimente, die sich jeweils im Hinblick auf die verwendete Datengrundlage unterscheiden, [lag] in erster Linie auf praktischen Fragestellungen u. a. der Korpusstichprobe, aber auch dem notwendigen Zeitbedarf und Arbeitsaufwand. […] Die Annotationsexperimente 3 und 4 wiederum dienten der Überarbeitung der Annotationsrichtlinien, um sie auf die Anforderungen der Anwendungsszenarien der vorliegenden Arbeit zuzuschneiden, und der Überarbeitung des Auswahlkriteriums für die annotierte Stichprobe, um einen Datenbias zu verhindern. Schließlich wurde in beiden Experimenten zusätzlich das Inter-Annotator Agreement bestimmt.“

Die folgende Abbildung ist aus Cramer (ebd.: 136) entnommen und gibt einen Überblick über die von der Autorin durchgeführten Experimente.

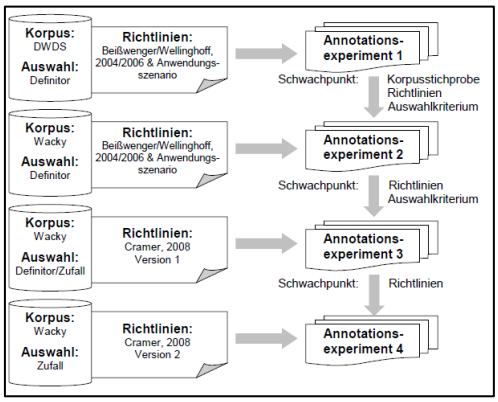


Abbildung 5: Annotationsexperimente nach Cramer (2011: 136)

Im Zuge ihrer Arbeiten kann Cramer – aufbauend auf den bereits existierenden Annotationsrichtlinien – das Annotationsverfahren schrittweise präzisieren und schließlich in

Experiment 4 einen Goldstandard annotieren. Dabei verwendet Cramer in den Experimenten 1-3 Stichproben aus deutschen Korpora, die Konkordanzen von „Definitorverben“, also lexikalischen Bestandteilen von Extraktionsmustern bzw. *knowledge patterns* darstellen. Im letzten Annotationsexperiment wird eine Zufallsstichprobe annotiert.

Die schließlich entwickelten Annotationsrichtlinien für Experiment 4 (ebd.: 238) definieren definitorische Textsegmente als „[…] Sätze, die als Ganzes oder in Teilen die Bedeutung von Begriffen erklären.“ Als Anwendungsszenario wird die Ergänzung einer lexikographischen (d. h. allgemeinsprachlichen) Ressource mit korpusbasierten Alltagsdefinitionen genannt. Des Weiteren werden Definiendum und Definiens als zwingende Bestandteile definitorischer Aussagen eingeführt, Möglichkeiten ihrer linguistischen Realisierung werden anhand von Beispielen erläutert. Die durch Definitionen gemäß der von Cramer entwickelten Typologie vermittelten Informationstypen werden anhand von Beispielen dargestellt. Strukturelle Typen von Definitionen (z. B. aristotelische oder parenthetische Definitionen) werden an Beispielen erläutert und typische Formulierungsmuster bzw. *knowledge patterns* werden aufgezählt und Informationstypen zugeordnet. Anders als Walter (2010) legt Cramer (2011: 253, Hervorhebungen im Original) dabei Wert auf eine einzelsatzbezogene Annotation von Definitionen:

> „Zunächst einmal gilt, dass **jeder Satz separat** im Hinblick auf die Frage analysiert wird, ob eine Definition vorliegt oder nicht. Dann stellt sich die Frage, ob das Textsegment auch ohne Kontext zentrale Informationen über ein Definiendum enthält, dass heißt [sic!], **ob die Information zum Definiendum lokal ungebunden ist. Lokal ungebunden bedeutet hier, die Information über das Definiendum muss unabhängig von dem Text in dem es vorkommt** [sic!] **gelten können, also eine vom Text unabhängige Information darstellen.** Ist das der Fall, muss das Segment als Definition annotiert werden.“

Die Explizitheit des Definiendums wird als weiteres wichtiges Entscheidungskriterium genannt (ebd.: 254). Anaphorische Elemente (z. B. inhaltliche Rückbezüge mit Hilfe von Demonstrativpronomina) und Eigennamen werden explizit als Negativmerkmale genannt, d. h. Kandidaten mit entsprechenden Elementen werden nicht als Definition annotiert (ebd.: 254 f.). Die inhaltlichen Kriterien der Vollständigkeit und Korrektheit der durch die Definition vermittelten Informationen werden dagegen aus der konkreten Annotationsaufgabe ausgeklammert, da eine inhaltliche Prüfung im Allgemeinen – dies zeigt gerade die Studie Walters (2010) – Domänenwissen voraussetzt (Cramer 2011: 256). Entsprechende Bewertungen können mithin bestenfalls in einem weiteren Bearbeitungsschritt vorgenommen werden.

Den Annotationsexperimenten 3 und 4 vorgeschaltet war eine recht umfangreiche Probeannotation, deren Ergebnisse gemeinsam besprochen wurden (ebd.: 177 f.). Insgesamt wurden von sieben Annotatoren 30 Texte mit ca. 1500 Sätzen annotiert (ebd.). Die Autorin (ebd.) beobachtet im Verlauf dieser Experimente eine besonders gute Übereinstimmung der Annotatoren bei „terminologischen bzw. wissenschaftlichen“ Definitionen und vermutet, dass deren Identifikation möglicherweise leichter sei als die von Alltagsdefinitionen.

Interessanterweise – dies soll abschließend erwähnt werden – führt die Autorin (ebd.: 148 ff.) nicht nur Annotationsexperimente, sondern auch eine Fragebogenstudie durch, in der anonyme Versuchspersonen Definitionskandidaten hinsichtlich ihres definitorischen Status bewerten. Von dieser Studie erhofft sich die Autorin Hinweise darauf, wie Alltagsdefinitionen bzw. Kandidaten für Alltagsdefinitionen von Rezipienten bewertet werden. Den Grad der Übereinstimmung der Versuchspersonen im Hinblick auf ihre Beurteilungen bezeichnet Cramer (ebd.: 149) angesichts dessen, dass den Studienteilnehmern keinerlei Annotationsrichtlinien zur Verfügung standen, als

bemerkenswert und knüpft an diese Beobachtung die Vermutung, dass auch ungeschulte Personen – wahrscheinlich anhand linguistischer Signale – definitorische von generischen Aussagen mit einiger Sicherheit unterscheiden können (ebd.: 149 f.). Konkret scheinen die Respondenten insbesondere operationale und aristotelische Definitionen zu bevorzugen, emotionale oder subjektive Textbestandteile dagegen als Negativmerkmal zu betrachten (ebd.: 151). Nicht zuletzt analysiert Cramer (ebd.: 153 ff.) auch solche Definitionen, die in einem Online-Forum bzw. von Studenten zur Klärung bestimmter Begriffe formuliert wurden, und erhofft sich von dieser Untersuchung Einblicke in die Sichtweise von Definitionsproduzenten. Auch hier findet die Autorin (ebd.: 158) Präferenzen für bestimmte Definitionstypen (aristotelische, wortassoziative, operationale und Synonymdefinition). Insbesondere aristotelische Definitionstypen scheinen dabei, wie bereits angemerkt, als besonders hochwertig zu gelten (ebd.: 159 f.).

3.3.2.2 Evaluation manueller Annotationen

Im Rahmen empirischer Studien ist es üblich, die Qualität humanannotierter Daten mit Hilfe eines Maßes für den Grad der Übereinstimmung zwischen den unterschiedlichen Annotatoren zu messen. Hinter dieser Vorgehensweise steht die Annahme, dass übereinstimmende Annotationsergebnisse auf real existierende Sachverhalte bzw. die gestellte Annotationsanforderung verweisen und somit einen kognitiven Konsens in Bezug auf diese widerspiegeln. Weichen die Ergebnisse der Annotatoren dagegen zu stark voneinander ab, wird die Annotation als arbiträr und damit als nicht valide angesehen (Artstein/Poesio 2008: 556 f., Hervorhebungen im Original):

> „Researchers who wish to use hand-coded data — that is, data in which **items** are labeled with **categories**, whether to support an empirical claim or to develop and test a computational model — need to show that such data are **reliable**. The fundamental assumption […] is that data are reliable if coders can be shown to **agree** on the categories assigned to units to an extent determined by the purposes of the study […]. If different coders produce consistently similar results, then we can infer that they have internalized a similar understanding of the annotation guidelines, and we can expect them to perform consistently under this understanding. Reliability is thus a prerequisite for demonstrating the **validity** of the coding scheme — that is, to show that the coding scheme captures the 'truth' of the phenomenon being studied […]."

Der Grad der Übereinstimmung in der Annotation der präsentierten Daten wird üblicherweise als *inter-annotator agreement* (IAA) bezeichnet. Formeln zur Berechnung des Inter-Annotator-Agreement haben die in 7) gezeigte Form (ebd.: 559):

7) $$IAA = \frac{\mathrm{Ao}-\mathrm{Ae}}{1-\mathrm{Ae}}$$

A_o bezeichnet dabei die empirisch beobachtete Übereinstimmung, d. h. den Anteil der Daten, der übereinstimmend einer der zur Verfügung stehenden Kategorien (im hier beschriebenen Fall: valide vs. invalide KRC-Kandidaten) zugeordnet wurde. A_e bezeichnet dagegen die bei zufälliger Annotation erwartete Übereinstimmung, die anhand der Annotationsdistributionen der Annotatoren berechnet wird. Aus diesen Angaben wird mit Hilfe der in 7) dargestellten Formel errechnet, wie weit die tatsächlich beobachtete Übereinstimmung über eine bloß zufällige Übereinstimmung hinausgeht (ebd.):

> „The value 1 − Ae will then measure how much agreement over and above chance is attainable; the value Ao − Ae will tell us how much agreement beyond chance was actually found. The ratio between

Ao−Ae and 1−Ae will then tell us which proportion of the possible agreement beyond chance was actually observed."

Üblicherweise werden die so errechneten Werte mit Hilfe einer von Landis und Koch (1977, zitiert nach Artstein/Poesio 2008: 576) entwickelten Intervall-Skala qualitativ bewertet. Allerdings handelt es sich hierbei um mehr oder weniger willkürliche Festlegungen (Landis/Koch 1977: 165), so dass beispielsweise Artstein und Poesio resümieren (2008: 576): „Unfortunately, deciding what counts as an adequate level of agreement for a specific purpose is still little more than a black art […]."
Eine alternative Möglichkeit für die Bestimmung der Güte der Annotation ist die Verwendung der Maße Präzision *(precision)* und Abdeckung (*recall*), aus denen sich das sogenannte f-Maß berechnen lässt[36]. Diese Maße sind insbesondere im Bereich des *information retrieval* gängig und wie folgt definiert (Manning/Schütze 1999: 267 ff.)

8) $precision = \frac{tp}{tp+fp}$

tp (true positives) bezeichnet in diesem Fall die Anzahl derjenigen gesuchten Elemente (in unserem Fall wissenshaltige Kontexte) bezeichnet, die mit Hilfe einer automatischen Extraktionsmethode in einer Datensammlung identifiziert wurden. *fp (false positives)* sind diejenigen Elemente, die fälschlicherweise als der Zielklasse zugehörige Elemente ebenfalls extrahiert wurden. *Precision* berechnet also den Anteil der gesuchten Elemente am Gesamtoutput eines Retrieval-Systems.

9) $recall = \frac{tp}{tp+fn}$

fn (false negatives) bezeichnet hier die Anzahl derjenigen gesuchten Elemente, die fälschlicherweise nicht extrahiert wurden. *Recall* ist also ein Maß für den Grad der Abdeckung, den ein Retrieval-System erreicht.

10) $f = (1+\beta^2)\,\frac{pr}{\beta^2\,p+r}$

β ist dabei ein zu wählender Gewichtungsfaktor und *P* und *R* beziehen sich jeweils auf Präzision und Abdeckung. Das f-Maß ist also ein Maß für die Gesamtperformanz des Systems. Bei Wahl des Gewichtungsfaktors β = 1 ergibt sich das f-Maß als harmonischer Mittelwert aus Präzision und Abdeckung (Walter 2010: 182 f.).

Werden diese Evaluationsmaße auf eine Annotationsaufgabe angewandt, wird angenommen, dass die erwartete Übereinstimmung $A_e = 0$ beträgt. Diese Annahme steht im Gegensatz zu der in 7) dargestellten Formel, in der sich A_e aus der Verteilung der zu differenzierenden Klassen im annotierten Datensatz ergibt. Praktisch werden die Klassenentscheidungen eines Annotators als Goldstandard betrachtet, an denen die Annotationen der verbleibenden Annotatoren mit Hilfe vorstehender Formeln gemessen werden. Die Annahme einer erwarteten Übereinstimmung $A_e = 0$ kann beispielsweise im Falle sehr komplexer Annotationsaufgaben gerechtfertigt sein.
Was die Wahl zwischen diesen unterschiedlichen Evaluationsansätzen betrifft, merken Sachs und Hedderich (2009[13]: 626) an, dass das am häufigsten angewandte IAA-Maß κ durch die Verteilung der Randsummen in der Kontingenztabelle (ebd.: 555, siehe auch die Beispiele für Vierfeldertafeln im

[36] Persönliche Mitteilung von Kevin Bretonnel Cohen, Universität Colorado, Denver, USA, http://compbio.ucdenver.edu/Hunter_lab/Cohen/index.shtml, vom 23.9.2013.

nächsten Kapitel) und damit letztlich durch die Verteilung der Klassen im annotierten Datensatz beeinflusst wird. Ergebnisse unterschiedlicher Annotationsstudien seien mithin nicht miteinander vergleichbar – eine sehr starke Einschränkung. Bei der Verwendung von Präzision, Abdeckung und f-Maß tritt diese Einschränkung nicht auf, da keine Informationen über negative Beispiele vorliegen müssen (vgl. beispielsweise Abschnitt 10.5.2 des GATE User Guides: http://gate.ac.uk/sale/tao/splitch10.html#x14-29000010.5, Cunningham et al. 2011, Zugriff am 28.03.2014). Diese Evaluationsmaße für das *inter-annotator agreement* sind deshalb insbesondere in den computerlinguistischen Teildisziplinen *information extraction* und *information retrieval* besonders beliebt.

3.3.3 Zusammenfassung

Die Annotation wissenshaltiger Kontexte ist, wie dargestellt, aus zweierlei Gründen von Interesse: Einerseits wird die Leistung computerlinguistischer Systeme in vielen Bereichen an entsprechenden menschlichen Leistungen gemessen, wobei hinter dieser Vorgehensweise der Gedanke steht, dass die menschliche Leistung einen sogenannten *benchmark*, d. h. gleichsam eine Obergrenze des technisch Erreichbaren darstellt. Andererseits können Annotationsstudien auch darüber Aufschluss geben, welche Arten von Aussagen von potentiellen Rezipienten überhaupt als definitorisch wahrgenommen werden und inwieweit diesbezüglich zwischen den Rezipienten von KRC-Kandidaten überhaupt ein Konsens besteht.

Die Analyse der verfügbaren Literatur hat allerdings in Bezug auf beide Aspekte ernüchternde Ergebnisse erbracht. Sowohl auf theoretischer als auch auf empirischer Grundlage argumentieren mehrere Autoren dahingehend, dass die Annotation definitorischer Aussagen bzw. wissenshaltiger Kontexte in mehrerlei Hinsicht problematisch sei, da sie auf komplexen Zusammenhängen bzw. Merkmalsbündeln beruhe. Nicht verwunderlich angesichts dieser Schwierigkeiten ist die Tatsache, dass in zahlreichen Studien ein Annotationsprozess nicht oder nur am Rande thematisiert und auf vereinfachte Modelle definitorischer Aussagen rekurriert wird, denn auch außerordentlich skrupulös vorgenommene Annotationen unterliegen – dies wird wohl auch vorliegende Arbeit kaum vermeiden können – angesichts der Komplexität der Annotationsaufgabe offensichtlicher Kritik und Revision. Ein negativer Nebeneffekt des weitgehenden Fehlens systematischer Annotationsstudien ist dabei jedoch die Nichtvergleichbarkeit von Extraktions- oder anderen empirischen Ergebnissen.

Ergebnisse umfangreicherer Annotationsstudien werden von Walter (2010) und Cramer (2011) vorgelegt, wobei zunächst auffällt, dass beide Autoren im Vergleich zu dem in Kapitel 2 entwickelten Begriff des wissenshaltigen Kontexts sowie den dort diskutierten Typologien mit – zumindest in den Annotationsrichtlinien – recht unspezifischen Definitionsbegriffen arbeiten. Andererseits vermittelt insbesondere Cramer (2011) ihren Annotatoren detaillierte Informationen zu linguistischen Erscheinungsformen definitorischer Aussagen, obgleich sie beobachtet, dass neben den sowohl in der Literatur als auch in der Alltagspraxis scheinbar präferierten Typen zahlreiche weitere Definitionsformen existieren. Gegen ein gewisses Maß an Zirkelhaftigkeit – den Annotatoren werden konkrete linguistische Merkmale oder *knowledge patterns* als Definitionssignale genannt und diese werden letztlich wieder als definitionstypische „Textsortenmerkmale“ bezeichnet – ist eine solche Vorgehensweise freilich nicht hundertprozentig gefeit. Umfassende Informationen über die Rezipienten- oder gar die Übersetzersicht auf definitorische Aussagen können auch diese Studien letztlich nicht liefern. Auch aus rein quantitativer Sicht schließlich demonstrieren die zitierten Studien die sich bereits im Theoretischen abzeichnende Schwierigkeit der endgültigen Unterscheidung der Klasse der definitorischen Aussagen von derjenigen der generischen Aussagen: Keine der genannten

Studien – auch nicht Cramers Annotationsexperiment, dem immerhin eine in mehreren Schritten aufwendig ausgearbeitete und mit zahlreichen Beispielen versehene Annotationsrichtlinie zugrunde liegt – erreicht ein IAA, das auch nur annähernd im Bereich des von Artstein und Poesio als anzustrebend bezeichneten Übereinstimmungsniveaus (0,7-0,8, vgl. Artstein/Poesio 2008: 591) liegt. Der tatsächliche Status der von einigen Autoren hervorgehobenen linguistischen Merkmale definitorischer Aussagen bleibt mithin zweifelhaft, d. h allein auf Grundlage der bislang geleisteten Literaturübersicht kann nicht entschiedenen werden, ob diese Merkmale tatsächlich eine Trennung der beiden Klassen definitorisch und generisch ermöglichen. Höhere κ-Übereinstimmungswerte scheinen vor diesem Hintergrund nur unter Zuhilfenahme ausführlicherer Annotationsrichtlinien erreichbar. Wie dann allerdings sichergestellt werden soll, dass das mittels dieser Richtlinien erfasste theoretische Konstrukt auch tatsächlich den praktischen Bedürfnissen der Nutzer definitorischer und erklärender Aussagen gerecht wird, ist fraglich. Hier sind eingehendere Studien erforderlich. Als weitere wichtige Unterscheidung kann auf Grundlage der Diskussion von in der Literatur beschriebenen Extraktionsexperimenten festgehalten werden, dass unterschiedlichste Typen definitorischer Aussagen in unterschiedlichen Settings – von Fachgebietsexperten oder aber Laien – annotiert werden.

3.4 Extraktion wissenshaltiger Kontexte

Nicht zuletzt ist in diesem Kapitel auch auf den gegenwärtigen Forschungsstand im Bereich der Extraktion definitorischer Aussagen einzugehen, da im weiteren Verlauf dieser Arbeit eigene Extraktionsexperimente konzipiert und durchgeführt werden sollen. Die zu betrachtenden Forschungsarbeiten stammen dabei nicht nur aus der Terminologieforschung (mit den Teilbereichen Terminographie und *ontology learning*), sondern auch aus der computerlinguistischen Teildisziplin der Informationsextraktion. Aus typologischer Sicht sind mehrere Ansätze zu unterscheiden: Wurden für die automatische Extraktion definitorischer Aussagen in der Vergangenheit insbesondere symbolische, musterbasierte Verfahren eingesetzt, werden diese in der letzten Zeit mit Methoden des maschinellen Lernens (ML) zu hybriden Herangehensweisen kombiniert. Nicht zuletzt wurden auch Extraktionsmethoden entwickelt, die ausschließlich auf Techniken des maschinellen Lernens beruhen.

3.4.1 Symbolische Verfahren

Terminologische Arbeiten zur Extraktion wissenshaltiger Kontexte stützen sich – anders als Arbeiten aus den computerlinguistischen Nachbardisziplinen – häufig noch sehr weitgehend auf musterbasierte Ansätze. Dies mag an dem Bemühen liegen, Kontexte zu extrahieren, die jeweils ganz bestimmte semantische Relationen verbalisieren, also genuin terminographischen Anforderungen und Arbeitsszenarien sowie möglicherweise dem zumindest auf theoretischer Ebene mitunter komplexeren und vielschichtigeren Herangehen an den Untersuchungsgegenstand liegen. Was hiermit konkret gemeint ist, stellt Meyer (2001: 289, Hervorhebungen im Original) pointiert dar. Sie geht in ihrer Arbeit von einem terminusorientierten Ansatz (*term-oriented approach*) aus, der sich dadurch auszeichnet, dass Termini bereits bekannt sind, zu ihnen aber noch semantisches Wissen erworben werden muss:

> „According to this scenario, the terminographer has a term in mind, and asks the system to search the corpus for those contexts that indicate a particular conceptual relation for this term. For example, 'find

> me functions for the term *compost*'. The system would provide a KWIC display of these contexts, allowing the user to manipulate the output in ways that have become standard in modern concordance tools […]."

Eine frühe terminologische Arbeit mit ähnlicher Ausrichtung ist der Aufsatz von Condamines und Rebeyrolle (1998), die eine Methode für den Aufbau einer terminologischen Wissensbank (*terminological knowledge base*) beschreiben. Aus einem französischen Korpus werden dafür mit Hilfe eines Termextraktors Termkandidaten extrahiert und nach lexikalischen Kriterien gefiltert. Alsdann werden Termtuplen mit linguistischen Suchmustern – einfachen Strings – kombiniert, um Kontexte zu extrahieren, auf deren Grundlage definitorische Muster spezifiziert werden können (vgl. auch die untenstehenden Ausführungen zum Thema *bootstrapping*). Diese Muster basieren auf Layout-Informationen ebenso wie auf lexikalischen und wortartenbezogenen sowie syntaktischen Angaben. Eine quantitative Evaluation der Ergebnisse findet nicht statt. In einem Artikel von 2001 (Condamines/Rebeyrolle 2001) präsentieren die Autorinnen eine leicht weiterentwickelte Variante dieser Methode, die in untenstehendem Abschnitt zum Thema *bootstrapping* kurz diskutiert wird.

Muresan und Klavans (2002: 231) extrahieren englische Alltagsdefinitionen für den Aufbau von Wörterbüchern. Auf Grundlage getaggter und gechunkter Texte zum Thema Medizin werden zunächst *knowledge patterns* erkannt (ebd.: 232). Alsdann setzt eine grammatische Komponente auf dem Output eines statistischen Parsers auf. Die Autoren erreichen auf diese Weise eine Präzision von 87% und eine Abdeckung von 75% (ebd.: 232). Des Weiteren führen Muresan und Klavans ein Qualitäts-Ranking für die automatisch extrahierten Definitionen im Vergleich zu Definitionen aus anderen Ressourcen durch, um zu bestimmen, wie acht Testpersonen (keine Fachgebietsexperten) die Definitionen beurteilen. Die Bewertung der Definitionen nach den zwei Kriterien Lesbarkeit und Nützlichkeit ergibt, dass die aus Texten extrahierten Definitionen ihren Vergleichsobjekten gegenüber überlegen sind (ebd.: 233). Eine Korrelationsstudie unterstützt dieses Ergebnis. Um die Perspektive von Fachgebietsexperten auf die Extraktionsergebnisse zu untersuchen, werden alsdann 15 medizinische Experten gebeten, die extrahierten Definitionen nach den Aspekten Genauigkeit und Vollständigkeit zu charakterisieren. Im Schnitt erhalten die Definitionen eine Genauigkeit von 5,38 und eine Vollständigkeit von 5,87 (beide Werte auf einer Skala von 1 – sehr schlecht – bis 7 – sehr gut). Allerdings werden für beide Evaluationen nur 15 Termini und die entsprechenden Definitionen herangezogen (ebd.).

Oberholzer (2002: 23 ff.) stellt anhand von Ergebnissen einer kleinen Umfrage unter Fachübersetzern der Bank „Credit Suisse" die Wichtigkeit definitorischer Angaben für die übersetzungsbezogene Terminologiearbeit heraus. Im experimentellen Teil ihrer Arbeit werden alsdann Extraktionsmuster für Inhalts- und Umfangsdefinitionen entwickelt (ebd.: 67). Als Testkorpus verwendet die Autorin ein deutsches Handbuch der Firma Siemens, das etwa 170.000 Wörter umfasst (ebd.: 68). Allerdings wurden nur in einem einzigen Kapitel des Handbuchs insgesamt 56 Definitionen manuell identifiziert (ebd.: 71). Manuell spezifizierte Suchmuster nutzen Lemma-Informationen zu *knowledge patterns*, Wortarten-*tags* sowie Informationen zu Nominal- und Präpositionalphrasen (ebd.: 68 ff.) ebenso wie eine Reihe von Zusatzbedingungen. Die Abfragen werden mit Hilfe der Programmiersprache Perl[37] durchgeführt (ebd.: 73 ff.), wobei nicht alle der von der Autorin als potentielle Suchmuster identifizierten Verben genutzt werden. Für die einzelnen Suchmuster erhält die Autorin Präzisionswerte zwischen 33-89% (ebd.: 77), wohingegen die musterbezogene Abdeckung zwischen 62-100% liegt (ebd.: 79). Abschließend beschreibt die Autorin noch Resultate einer weiteren manuellen Analyse von Begriffsdefinitionen, die interessante Ergebnisse erbringt, z. B. die Technik

[37] http://www.perl.org/, Zugriff am 21.03.2014.

des Definierens nicht nach aristotelischem Schema, sondern durch den Verbinhalt sowie die Verwendung von Parenthesen, typischen Präpositionen und Konjunktionen etc. (ebd.: 80 f.).
Malaisé et al. (2005: 24) verwenden für ihre Studie zwei morpho-syntaktisch annotierte Web-Korpora. Lexikalisch-syntaktische Muster werden auf Grundlage von Literaturstudien, der Analyse von Wörterbuchdefinitionen sowie der Ergebnisse einer ersten, explorativen Analyse der Untersuchungskorpora entwickelt (ebd.: 33). Da das Ziel dieser Arbeit nicht nur die Extraktion wissenshaltiger Kontexte, sondern die Anreicherung einer Ontologie ist, sollen zudem auch das Definiendum sowie Informationen zu zwischen diesem und anderen Termini bestehenden semantischen Relationen extrahiert werden. Für die Identifikation des Definiendums werden linguistische Heuristiken angewandt. Die Identifikation einer spezifischen semantischen Relation erfolgt dagegen anhand des einen Satz extrahierenden Musters (ebd.: 35 ff.). Da das Korpus in einem XML-Format vorliegt, können die Extraktionsmuster als XSLT-Regeln[38] implementiert werden. Auf ungesehenen Korpusdaten erreichen die Autoren (ebd.: 38 ff.) eine Präzision von 67% und eine Abdeckung von 39%.
Storrer und Wellinghoff (2006: 2373 f.) betrachten in ihrer Studie deutsche aristotelische Definitionen ebenso wie Wörterbuchdefinitionen. Diese Definitionen sollen Informationen über semantische Relationen zwischen Begriffen liefern, die wiederum in eine semantische Ressource eingepflegt werden sollen (ebd.: 2373). Die Autorinnen (ebd.: 2375) nutzen ein proprietäres Extraktionswerkzeug für ihren musterbasierten Ansatz und erreichen für ihre Suchmuster ebenfalls stark unterschiedliche Präzisionswerte. Informationen zu den linguistischen Datentypen, auf denen die Suche konkret basiert, werden nicht gegeben.
Del Gaudio und Branco (2007) arbeiten ebenso wie Przepiórkowski et al. (2007), Borg (2009) und Westerhout (2010) nicht in einem terminologischen Kontext, in dessen Rahmen terminographisch relevantes Wissen extrahiert oder semantische Ressourcen aufgebaut werden sollen, sondern innerhalb eines *eLearning*-Projekts. Ähnlich wie Borg (2009) und Westerhout (2010) konzentrieren sich die Autoren auf die Extraktion von Kopula-Definitionen mit dem Verb „sein“, Definitionen mit anderen Konnektorverben sowie Definitionen mit spezifischer Interpunktion (Del Gaudio/Branco 2007: 660), die untersuchte Sprache ist Portugiesisch. Das Untersuchungskorpus deckt unterschiedliche Fachgebiete ab, umfasst 274.000 Wörter und wurde mophosyntaktisch annotiert (ebd.: 661). Manuell wurden insgesamt 651 Definitionen annotiert. Die Extraktionsregeln werden ebenso wie in der Arbeit von Westerhout (2010) als XML-basierte *rewrite*-Regeln implementiert. Insgesamt erreichen die Autoren (Del Gaudio/Branco 2007: 666) auf ihrem Testkorpus mit dieser Methode eine Präzision von 14% und eine Abdeckung von 86%. Die Autoren (ebd.) merken abschließend an, dass Extraktionsergebnisse auch textabhängig seien:

> „Documents with a clear educational purpose, like those from IS and IT sub-corpora, are more formal in the structure and are more directed towards explaining concepts, many times via the presentation of the associated definitions. On the other hand, documents with a less educative purpose present less explicit definitions and for this reason it is more difficult to extract definitory contexts from them using basic patterns.“

[38] XSLT ist Teil von XSL, der Extensible Stylesheet Language, und stellt eine Programmiersprache dar, mit deren Hilfe XML-Dokumente transformiert werden können. Es existiert eine offizielle Empfehlung des *World Wide Web Consortium* (W3C) für die Spezifikation von XSLT: http://www.w3.org/TR/xslt.html, Zugriff am 25.07.2014.

Damit bestätigen die Autoren Aussagen Meyers (2001: 299). Nicht zuletzt wird – wie auch in anderen Arbeiten – die Schwierigkeit der Evaluation von Extraktionsergebnissen bemerkt, die auf die Komplexität der Kategorie „Definition“ zurückgeht (Del Gaudio/Branco 2007: 667).

Przepiórkowski et al. (2007: 43) verwenden morphosyntaktisch annotierte Korpora in einem XML-Format, die untersuchten Sprachen sind Bulgarisch, Tschechisch und Polnisch (ebd.: 44 ff.). Die Extraktionsregeln werden auch hier als XML-basierte *rewrite*-Regeln implementiert. Als Testset werden für jede Sprache etwa 150 bislang ungesehene Definitionen verwendet, wobei die Definitionen auch mehrere Sätze umfassen können. Ebenso wie später bei Jin et al. (2013) wird die Evaluation auf der Ebene einzelner *tokens* durchgeführt (Przepiórkowski et al. 2007: 46). Aufgrund dieser Evaluation werden für die einzelnen Sprachen Präzisionswerte zwischen 15-21% und Abdeckungswerte zwischen 2-41% erreicht (ebd.: 47). Eine zweite Evaluation lediglich auf Grundlage einer partiellen Übereinstimmung mit manuell annotierten Definitionen ergibt eine Präzision von 23% und eine Abdeckung zwischen 9-46% (ebd.: 47). Eine Ursache für Performanzprobleme sind nach Aussage der Autoren (ebd.: 48) die für slawische Sprachen typischen stark ausgeprägten morphologischen Paradigmen, die *tagging*- und daraus resultierende Verarbeitungsfehler auf höheren Ebenen nach sich zögen. Eine weitere Schwierigkeit bestehe in der freien Wortfolge slawischer Sprachen, zu deren Abdeckung umfangreichere Regelsätze zu implementieren seien.

Sierra et al. (2008) präsentieren eine Studie zur Extraktion wissenshaltiger Kontexte aus einem spanischen Korpus, das eine Reihe unterschiedlicher Fachgebiete abdeckt. Zunächst werden für eine Reihe ausgewählter Verben maximal 300 mit Wortarten-*tags* versehene Korpus-Konkordanzen extrahiert (ebd.: 86). Das *knowledge pattern* wird alsdann ebenfalls als solches getaggt (ebd.: 87), mit Hilfe einiger linguistischer Regeln wird die Ergebnisliste bereinigt (ebd.: 88 f.). Für die Identifikation der Kontext-Bestandteile Definiendum, *knowledge pattern* und Definiens wird mit dem Entscheidungsbaum-Verfahren[39] sodann eine populäre Methode des maschinellen Lernens eingesetzt. Die Autoren erreichen eine durchschnittliche Präzision (Mittelwert über den errechneten Präzisionswerten der einzelnen Muster) von 61%. Die Abdeckung wird, wie bereits dargestellt, nicht angegeben.

Die Analyse der Literatur zu musterbasierten, symbolischen Verfahren verweist zunächst auf eine wesentliche Unterscheidung. So beruhen einige der diskutierten Methoden auf einer vorhergehenden Termerkennung, es werden also definitorische Aussagen für konkrete *Zieltermini* gesucht (in Meyers (2001) Terminologie ein *term-oriented approach*). Andere Ansätze streben dagegen danach, *alle* in einem Korpus vorhandenen Definitionen oder wissenshaltigen Kontexte zu identifizieren. Die Analyse zeigt außerdem, dass musterbasierte Ansätze in der Regel auf relativ einfachen Verarbeitungsmethoden und Beobachtungen auf der Ebene der linguistischen Oberflächenstruktur (Lemma, Wortarten-*tag*, evtl. morphologische Informationen und *chunks*/Phrasenstrukturinformationen) beruhen. Häufig wird eine einfache Beziehung zwischen einem Extraktionsmuster und semantischen Relationen angenommen – eine Vermutung, die durch die Extraktionsergebnisse meist nicht vollumfänglich bestätigt werden kann. Zu fragen ist mithin, ob durch komplexere Suchmuster nicht bessere Ergebnisse erreicht werden können. So differenziert Waterman (1998: 149 f.) zwischen kategorischen (auf semantische und funktionale Klassen bezogenen), morphologischen und strukturellen (auf die Wortfolge bezogenen) Merkmalen

[39] Das Entscheidungsbaum-Lernverfahren besteht im Erlernen einer Schrittfolge von Entscheidungen auf Grundlage der vorhandenen Merkmalsmenge, in dessen Ergebnis Instanzen des zu untersuchenden Datensatzes klassifiziert werden können. Dabei werden die Merkmale in der Reihenfolge ihrer Eignung für die Bewältigung der zu lösenden Klassifikationsaufgabe für das Erlernen von Regeln selektiert (vgl. Witten et al. 2011[3]: 99 ff.).

lexikalischer Muster. Meyer (2001: 290) dagegen unterscheidet lexikalische, grammatische und paralinguistische Aspekte der Musterbildung. Walter (2010: 160 ff.) nennt als grundlegende Mustertypen sequenzbasierte Muster, die auf die Elemente *string* (aufeinander folgende Buchstabenfolgen), Lemma (Sequenzen bestimmter Lemmata), Wortart (Sequenzen von Wortart-*tags*) oder Satzposition (Vor- oder Nachstellung von Definitionsbestandteilen) rekurrieren, und dependenzbasierte Muster, die syntaktische Abhängigkeitsbeziehungen beschreiben. Dependenzbasierte Muster sind dabei insbesondere zur Spezifikation der Argumentstruktur von häufig in relevanten Sätzen auftretenden Prädikaten geeignet[40].

Neben diesem möglicherweise in empirischen Untersuchungen noch nicht vollständig ausgereizten Potential birgt die Verwendung sprachlicher Muster aber auch eine Reihe von Schwierigkeiten. Zwar nennt Hearsts (1992: 540) Aufsatz neben Mustern für die Hyperonymie-/Hypernymie-Relation auch einige erste Qualitätsanforderungen für *knowledge patterns*, und zwar:

- häufiges Auftreten des Musters auch in unterschiedlichen Texttypen und –genres,
- hohe Zuverlässigkeit: markiert (fast) immer die interessierende Relation,
- erfordert nur geringen Vorverarbeitungs-Aufwand.

Allerdings ist die Formulierung von Extraktionsmustern ein nur schwer kontrollierbarer und insgesamt wenig transparenter Prozess. So kritisiert Waterman (1998: 155):

> „The hardest and most time-consuming task involved is certainly the acquisition of patterns, which is still done primarily by tedious hand analysis. [...] Although the key templates are originally created by a researcher doing a careful reading, the correspondence between text segments and the key templates has not been recorded, making the process error prone and leaving the text open for reinterpretation."

Auf die Ausführungen Meyers (2001) und Condamines' (2002) zu Schwierigkeiten im Zusammenhang mit der Verwendung linguistischer Extraktionsmuster war bereits in Abschnitt 3.2.3 hingewiesen worden. Ob und inwieweit die von Hearst formulierten Qualitätsanforderungen an Extraktionsmuster tatsächlich realistisch umsetzbar sind, ist folglich offen.

3.4.2 Maschinelles Lernen

Verfahren des maschinellen Lernens (ML) werden eingesetzt, um statistische oder regelbasierte Modelle für die automatische Klassifikation unbekannten Textmaterials zu erstellen, oder um deterministische Regeln für die Extraktion definitorischer Aussagen vollautomatisch oder mit nur minimaler menschlicher Unterstützung zu generieren. Entsprechend können ML-Verfahren in Klassifikationsverfahren und Methoden des automatisierten Erwerbs linguistischer Regeln unterteilt werden.

3.4.2.1 Klassifikationsverfahren

Fahmi und Bouma (2006) testen drei unterschiedliche Algorithmen für das „überwachte" maschinelle Lernen, und zwar *Naive Bayes*, *Maximum Entropy* (ME) und Support-Vektor-Maschinen (Fahmi/Bouma 2006: 68; eine leicht verständliche Einführung in wesentliche Algorithmen geben beispielsweise Witten et al. 2011[3], eine stärker mathematisch formalisierte Darstellung findet sich in

40 Walters (2010) Studie, in der tatsächlich mittels syntaktischer Muster ein besseres Gesamtergebnis erzielt wird, wird im Zusammenhang mit hybriden Verfahren diskutiert.

Murphy 2012). Das Grundprinzip der Anwendung „überwachter“ Lernverfahren besteht dabei darin, dass auf Grundlage von Trainingsdaten, die mittels einer vorab definierten Menge von Merkmalen (*features*) beschrieben werden, zunächst ein Modell erlernt wird, mit dessen Hilfe die Trainingsdaten im Hinblick auf die interessierende Fragestellung möglichst optimal charakterisiert werden können. Dieses Modell kann alsdann auf ungesehene Testdaten angewandt werden, um beispielsweise Definitionen von generischen Sätzen zu unterscheiden. Dieser Schritt wird als Klassifikation bezeichnet und besteht in der Zuordnung von Instanzen des zu untersuchenden Datensatzes zu einer der vorab definierten Klassen, d. h. hier entweder zur Klasse der definitorischen oder aber derjenigen der generischen Aussagen. Für ihre Experimente verwenden Fahmi und Bouma die folgenden *features* (in einer recht eigenwilligen Zuordnung zu den linguistischen Kategorien Text, Dokument und Syntax) (ebd.: 67):

- textbasierte Merkmale, z. B. alle Wörter in einem Satz (*bag of words*), Wortstämme, Zweiwortfolgen (*bigrams*), Interpunktion,
- dokumentbezogene Merkmale, z. B. die Position des Satzes im Dokument (der bereits diskutierten Annahme der Konzentration von Definitionen in bestimmten Textabschnitten folgend),
- syntaktische Merkmale, z. B. die Position des Subjekts in einem Satz und der Typ der Determinatoren vor Subjekt und Objekt,
- Eigennamen, d. h. Wortarten-*tags*, die auf real existierende Organisationen, Orte oder Personen verweisen, da die Autoren, wie ebenfalls bereits diskutiert, davon ausgehen, dass Definitionen insbesondere in Subjektposition mehr Eigennamen enthalten.

Die Performanz dieser Verfahren wird auf einem Datensatz getestet, der aus 1336 Definitionen und 963 generischen Sätzen besteht (womit im Vergleich zur Extraktion definitorischer Aussagen aus authentischen Texten eine unrealistische Versuchsanordnung gegeben ist, da kaum damit gerechnet werden kann, dass definitorische Aussagen in Texten mehr als 50% der Satzmenge ausmachen) (ebd.: 69). Als Evaluationsmaß wird u. a. die Korrektheit der Klassifikation (*accuracy*) geschätzt, d. h. die Anzahl derjenigen Sätze, die von dem jeweiligen Lernverfahren korrekt als definitorisch oder generisch klassifiziert wurden (vgl. Borg 2009: 8). Das beste Ergebnis (92,21% korrekte Klassifizierungen) erzielt der ME-Klassifizierer auf Grundlage der Merkmale *bag of words, bigrams* und syntaktische Information nach Anwendung eines simplen lexikalischen Filters.

Navigli und Velardi (2010) verwenden sogenannte *word-class lattices*, um Sätze in Texten als definitorisch oder generisch zu klassifizieren. *Word-class lattices* sind gerichtete azyklische Graphen (ebd.: 1318), mit deren Hilfe strukturelle Eigenschaften von Definitionen verallgemeinert und repräsentiert werden sollen. Die Autoren betrachten lediglich aristotelische Definitionen (ebd.: 1320 ff.). Als Trainingsmaterial werden Sätze verwendet, in denen Definiendum und *genus proximum* sowie das beide Elemente verbindende Verb und weitere Satzbestandteile manuell annotiert wurden. Alsdann werden die untersuchten Sätze dahingehend vereinfacht, dass selten auftretende Wörter durch einen Platzhalter (*) ersetzt werden. Diese Verallgemeinerung betrifft im Wesentlichen die Struktur „Definiendum – VERB – *genus proximum*“ und damit die auch von Borg (2009) im Wesentlichen betrachtete Struktur. Weitere Satzbestandteile (z. B. möglicherweise angeschlossene Relativsätze) werden in der Folge nicht mehr betrachtet (Navigli/Velardi 2010: 1321). In einem weiteren Schritt werden die so erhaltenen verallgemeinerten Sätze nach Ähnlichkeit gruppiert (*geclustert*). Die *word-class lattices* werden dann konstruiert, indem die korrespondierenden Bestandteile der Sätze in einem Satzcluster miteinander aligniert werden (ebd.: 1321 f.). Abbildung 6 ist der Arbeit von Navigli und Velardi entnommen und stellt eine *word-class lattice* dar.

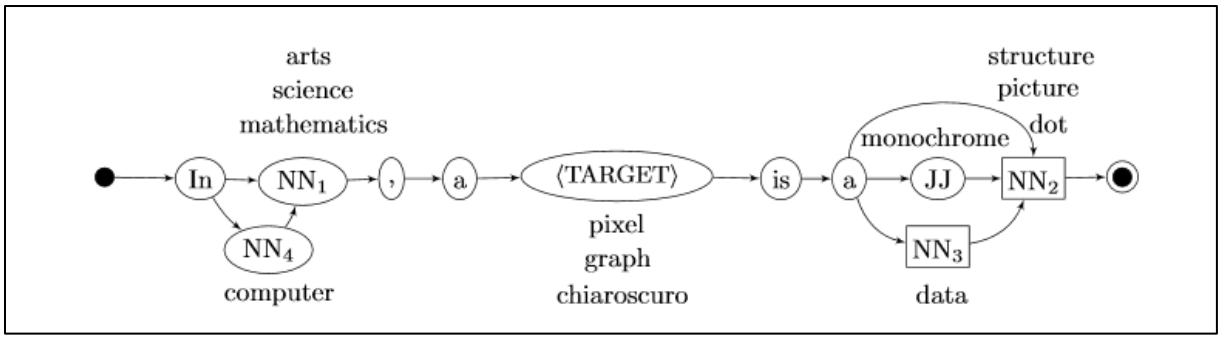


Abbildung 6: Word-class lattice für aristotelische Definitionen nach Navigli und Velardi (2010)

Mit Hilfe dieser Definitionsgraphen können im Ergebnis ungesehene Sätze klassifiziert werden, d. h. bei Übereinstimmung eines Definitionskandidaten mit einem der Definitionsgraphen wird der Satz als Definition klassifiziert, andernfalls als generischer Satz (ebd.: 1322). Die Autoren beschreiben zwei Klassifikationsexperimente. In einem ersten Experiment werden 1908 definitorische und 2711 generische Sätze aus der englischen Wikipedia als Testdaten ausgewählt (womit ebenfalls eine allzu optimistische Verteilung von Definitionen in den Testdaten im Hinblick auf ihre Häufigkeit in natürlichen Texten erzeugt wird) (ebd.: 1323). Die definitorischen Sätze entsprechen dabei den ersten Sätzen der Wikipedia-Artikel. Für ein zweites Experiment werden dagegen über 300.000 Sätze aus einem englischen Web-Korpus ausgewählt, die Okkurrenzen ausgewählter Termini enthalten. Für das erste Experiment erhalten Navigli und Velardi eine Präzision von nahezu 100% und eine Abdeckung von 42% (f-Maß = 0,59) (ebd.: 1324). Eine Variante ihres Algorithmus', in der *word-class lattices* nicht für ganze Sätze, sondern für die während der Annotation unterschiedenen Definitionsbestandteile erlernt werden, erreicht ein f-Maß von 0,75. Allerdings zeigen die Autoren auch, dass bereits bei einer Klassifikation aufgrund der verallgemeinerten Sätze mit Platzhaltern (*) ein f-Maß von 0,75 auf diesem Datensatz erreicht werden kann. Im zweiten Experiment wird eine Präzision von 98% erreicht, wohingegen die Abdeckung auf 39% geschätzt wird (die Variante der Klassifikation von Definitionsbestandteilen erreicht eine Präzision von 95% und eine Abdeckung von 57%, die Sätze mit Platzhaltern ergeben eine Präzision von 44% und eine Abdeckung von 64%).

Cramer (2011: 183 ff.) setzt auf Grundlage der von ihr annotierten Daten sowie der ermittelten linguistischen Merkmale von Definitionen mit Hilfe überwachter Verfahren des maschinellen Lernens ein einfaches Extraktionssystem um. Als *features* verwendet Cramer dabei lediglich die bereits beschriebenen lexikalischen Merkmale. Die von ihr getesteten ML-Algorithmen erbringen höchst unterschiedliche Ergebnisse, wobei jedoch alle eine *accuracy* von über 99% erreichen (angesichts von ca. 200 Definitionen in 37.500 zu klassifizierenden Sätzen nicht erstaunlich, da auch ein Algorithmus, der alle Sätze als generisch klassifiziert, eine solche Klassifikationsgenauigkeit erreichen wird). Als das mit diesem einfachen System beste errechnete f-Maß nennt Cramer (ebd.: 189 f.) den Wert 0,23 (37% Präzision und 17% Abdeckung), den sie mit einem Entscheidungsbaum-Algorithmus erreicht.

Jin et al. (2013) betrachten ebenso wie die vorgenannten Autoren – mit Ausnahme Cramers (2011) – ein eng gefasstes Spektrum aristotelischer englischer Definitionen (Jin et al. 2013: 782), die sie wissenschaftlichen Artikeln entnehmen, wobei die Klassifikation allerdings nicht auf Satz-, sondern auf *token*-Ebene, d. h. auf der Ebene einzelner Wörter durchgeführt wird. In diesem Szenario werden die Wörter in dem zu untersuchenden Datensatz dahingehend getaggt, ob sie Bestandteil einer Definition bzw. eines Definiendums sind oder nicht. Diese Herangehensweise wird als *sequence labelling* bezeichnet (ebd.: 780). Die Autoren verweisen ausdrücklich darauf, dass Wikipedia-Korpora aufgrund der stilistischen und formalen Konventionen, die das Verfassen von Wikipedia-Artikeln

regulieren, nicht für die Entwicklung generischer Extraktionssysteme geeignet sind (ebd.: 782): „[… W]e note that Wikipedia has stylistic conventions that make detection of definitions much easier than in the general case […]. This makes it unsuitable for training a general extraction system from scholarly text." Die Autoren verwenden folgende *features* (ebd.: 783):

- lexikalische Merkmale, z. B. Wortform, Wortstamm, Wortarten-*tag*,
- orthographische Merkmale, z. B. Groß- und Kleinschreibung,
- termbezogene Merkmale, z. B. die inverse Dokumenthäufigkeit (*inverse document frequency)* eines Wortes, die auch für die Berechnung von TF-IDF herangezogen wird,
- positionsbezogene Merkmale, z. B. den Namen des zugehörigen Textabschnitts,
- musterbezogene Merkmale, die das Auftreten von *knowledge patterns* beschreiben,
- syntaktische Merkmale, z. B. Informationen zum syntaktischen Kopf des gerade zu klassifizierenden Wortes sowie zum Abstand des Letzteren zur Wurzel des Syntaxbaumes und zur Art des Abhängigkeitspfades.

Im Vergleich zu einem Basis-Extraktionssystem, das lediglich lexikalische Merkmale nutzt, bewirkt das Hinzunehmen zusätzlicher *features* – mit Ausnahme der positionsbezogenen Merkmale – stets Qualitätsverbesserungen, so dass die Autoren letztendlich ein f-Maß von 0,49 (Präzision von 45% und Abdeckung von 54%) erreichen, wobei zu beachten ist, dass hier die Evaluation eben auf der Ebene des Einzelworts erfolgt (ebd.: 784 f.). Wird der eigentlichen Klassifikation ein Term-Identifikationsschritt vorgeschaltet, verbessert sich das f-Maß weiterhin bis auf 0,56. Jin et al. testen ihr System auch auf dem ersten der von Navigli und Velardi (2010) verwendeten Datensätze und evaluieren auf Satzebene. In diesem Experiment wird das von Navigli und Velardi erreichte f-Maß auf 0,85 gesteigert. Interessanterweise untersuchen Jin et al. (2013) auch einige Eigenschaften der Definitionen in ihrem Korpus. Demzufolge treten Definitionen zwar nicht nur am Beginn von Texten auf (was, wie bereits dargestellt, auch von Cramer 2011 bestritten wird), allerdings sind sie am Textbeginn häufiger anzutreffen als am Textende (was, mit einer gewissen Skepsis, ebenfalls von Cramer (ebd.: 185 ff.) bestätigt wird) (Jin et al. 2013: 787). Nicht zuletzt verweisen Jin et al. auf Grundlage ihrer Studie auch auf disziplinspezifische Trends, z. B. in Bezug auf die Anzahl neuer Definitionen pro Jahr oder besonders häufig definierte Termini.

3.4.2.2 Regelerwerb

Eine ausführliche Beschreibung eines recht elaborierten ML-Ansatzes, mit dessen Hilfe allein Definitionen extrahiert werden sollen, wird in der Arbeit von Borg (2009) vorgestellt. Die Autorin verwendet „genetische Programme" und „genetische Algorithmen", um Extraktionsregeln automatisch zu erlernen und ihre Bestandteile zu gewichten (ebd.: 34, Hervorhebungen im Original):

> „A Genetic Algorithm (GA) […] is a search technique which emulates natural evolution, attempting to search for an adequate solution to a problem by mimicking natural selection. By simulating a population of individuals represented as strings […] GAs try to evolve better solutions by selecting the best performing individuals (through the use of a *fitness function*), allowing them to survive and reproduce. This is done using two operations called *crossover* and *mutation*. […] This process mimicks the survival of the fittest, with the better performing individuals remaining within the population, and poor performing individuals being removed."
>
> „Genetic Programming (GP) […] is based on GAs, with the difference that the individuals are represented as programs (stored as trees), as opposed to strings as in GAs. […] The same types of operations are carried out in GPs, using sub-trees for creating new individuals in the crossover phase, and changing one randomly in the mutation phase."

Die Autorin arbeitet dabei auf drei unterschiedlichen Datensätzen, und zwar einem Trainings-, einem Test- sowie einem Evaluationsdatensatz, die jeweils 150 manuell annotierte Definitionen enthalten (ebd.: 62). Die Daten liegen in einem XML-Format vor. In früheren Arbeitsschritten wurden Extraktionsregeln, „[…] mainly POS sequences […]“ (ebd.), bestimmt. Auf Grundlage erster Extraktionsergebnisse unterscheidet die Autorin *is-a*-Definitionen, die eine Hyperonymie-Beziehung ausdrücken, Definitionen, die durch andere Verben als *to be* ausgedrückt werden, sowie solche Definitionen, in denen die Verbindung zwischen Definiendum und Definiens durch Interpunktion, Pronomina oder spezielle Mittel des Textlayouts realisiert wird (ebd.: 64 f.). Mit Hilfe von Extraktionsregeln erreicht die Autorin ein f-Maß von 0,33 für den Typus der Verb-Definitionen, für andere Definitionstypen dagegen noch schlechtere Resultate (ebd.: 67). Ausgehend von diesen Ergebnissen beschreibt Borg in ihrer Arbeit zwei Experimente. In einem ersten Experiment sollen Extraktionsregeln mit Hilfe eines genetischen Algorithmus‘ gewichtet werden (ebd.: 70):

> „From the experience gained through the manual crafting of grammar rules for definition extraction, we noticed that certain rules, or sub-parts, contain more specific or important information than other rules. This led to the idea of using a GA as a possible technique to learn the importance of features that can recognise definitions. This can be done by assigning weights to each feature and allowing the algorithm to adjust the weights according to the performance.”

In einem solchen Ansatz wird ein Definitionskandidat mit Hilfe einer Menge spezifischer Eigenschaften (*features*; in der Arbeit Borgs Extraktionsregeln, die auf Definitionskandidaten gematcht werden) beschrieben, die zur Errechnung eines Evaluationsmaßes, d. h. für die Entscheidung, ob ein Kandidat eine Definition ist oder nicht, linear miteinander kombiniert werden. Die Lernaufgabe besteht folglich im Erlernen der individuellen Feature-Gewichte für die lineare Kombination (ebd.: 70 f.). Als Grundlage für die Optimierung der selektierten Regelmenge (*fitness function*) werden dabei die bereits eingeführten Maße Präzision, Abdeckung und f-Maß verwendet (ebd.: 73 ff.). In einem zweiten Experiment schließlich sollen Extraktionsregeln mit Hilfe eines genetischen Programms vollautomatisch erlernt werden. Als *fitness function* kommen auch hier wieder die bereits vorgestellten Evaluationsmaße Präzision, Abdeckung und f-Maß zum Einsatz (ebd.: 81 ff.).

Im weiteren Verlauf ihrer Arbeit konzentriert sich die Autorin auf die Extraktion von *is-a*-Definitionen mit Hilfe eines genetischen Algorithmus‘, wobei die Definitionen in den untersuchten Daten mit einer Häufigkeit von 111 Definitionen gegenüber 21.112 generischen Sätzen vertreten sind (ebd.: 84). Die in diesem Experiment verwendeten *features* sind (ebd.: 85 f.):

- Satz enthält das Verb *to be*, gefolgt von einem Artikel,
- Satz enthält die *tag*-Sequenz „Fremdwort […] sein“, wobei die Reihenfolge dieser beiden Elemente variabel zu sein scheint,
- Satz enthält mindestens ein Possessivpronomen,
- Satz enthält ein Interpunktionszeichen (Bindestrich oder Komma) in der Mitte,
- Satz enthält einen vorher annotierten Zielterminus,
- Satz enthält typographische Hervorhebungen (Kursivschrift oder Fettdruck),
- Satz enthält einen *chunk*, der als Organisation, Person oder Ort annotiert ist.

Die Autorin testet unterschiedliche *settings* ihres Algorithmus‘ und erreicht ein maximales f-Maß von 0,57 (ebd.: 91 f.) als Ergebnis der *fitness function* für die beste gelernte Regel, kann allerdings zeigen, dass durch Modifikationen der fitness function f-Maß noch deutliche Veränderungen erzielt werden können (ebd.: 93 ff.). Die gelernten Gewichte ihrer *features* interpretiert Borg im Hinblick darauf, ob sie positive oder negative Signale für das Vorliegen einer Definition sind. Negative Gewichte, wie sie

für die Merkmale „Possessivpronomen“, „typographische Hervorhebung“ und „Organisations-*chunk*“ gelernt wurden, deuten entsprechend auf das Vorliegen eines generischen Satzes hin (ebd.: 97). Allerdings weist die Autorin (ebd.: 98) auch darauf hin, dass konkrete *feature*-Gewichtungen auch von den *settings* des Algorithmus‘ abhängen, so dass nur die *features* „typographische Hervorhebung“ und „Organisations-*chunk*“ eindeutig negativ seien – alle anderen Merkmale seien „[…] somewhat positive“ (ebd.: 98). Die Klassifikation von Definitionen im Trainingskorpus mit Hilfe der selektierten besten Extraktionsregel erweist sich jedoch als schwierig (ebd.: 99 ff.). Zwar führt die Autorin für dieses Experiment keine Performanzwerte im Sinne der eingeführten Evaluationsmaße an, die beschriebenen Probleme ähneln jedoch den andernorts in der Literatur (z. B. Walter 2010) genannten.

Mit Hilfe genetischer Programme erlernt Borg (ebd.: 105 ff.) schließlich Extraktionsregeln für unterschiedliche Definitionstypen. Für *is-a*-Definitionen erreicht die beste Extraktionsregel ein f-Maß von 0,28 (ebd.: 116), wobei im Wesentlichen Variationen der einfachen Extraktionsregel „Nomen – sein – Artikel – Nomen“ gelernt werden (ebd.: 117). Für diejenigen Definitionen, die durch andere Verben als die Kopula „sein“ (*to be*) ausgedrückt werden, werden auch morphologische Informationen (in Form von *tags*) sowie Informationen zum Vorliegen bestimmter Lexeme (Bausteine für *knowledge patterns*) für den Regelgenerierungsprozess in Betracht gezogen. Das maximal erreichte f-Maß liegt hier bei 0,21 (ebd.: 121). Schließlich werden auch Regeln für die Extraktion von Definitionen aufgrund von Interpunktion erlernt, wobei ein maximales f-Maß von 0,3 erzielt wird (ebd.: 122). Nicht zuletzt führt Borg (ebd.: 123 f.) auch Experimente durch, in denen der Suchraum für den Algorithmus erhöht wird, indem alle Definitionstypen miteinander vermischt werden, wobei wiederum im Wesentlichen Variationen der Regeln *is a* und *is called a* erlernt werden und ein maximales f-Maß von 0,22 erzielt wird. Schließlich erreicht Borg (ebd.: 128 f.) dadurch, dass sie die mit Hilfe von genetischen Programmen erzeugten Extraktionsregeln für *is-a*-Definitionen durch einen genetischen Algorithmus gewichten lässt, ein endgültiges f-Maß von 0,68 (100% Präzision und 51% Abdeckung bei der Verwendung der erlernten Regeln für die Klassifikation der Testsätze in definitorisch und generisch) für diesen Definitionstyp.

Ein weiterer Ansatz für die Generierung von Extraktionsregeln ist das *bootstrapping*, das insbesondere in Arbeiten zur Informationsextraktion eingesetzt, aber auch in terminologischen Studien diskutiert wird. Die Grundidee des *bootstrapping* wird bereits von Hearst (1992: 541) beschrieben:

- Bildung von *seed*-Termpaaren (Tuplen), zwischen denen die semantische Zielrelation besteht,
- Studie von Korpus-Okkurrenzen solcher Tuplen,
- Verallgemeinerung der in diesen Okkurrenzen gefundenen Gemeinsamkeiten zu sprachlichen Mustern,
- Nutzung der aufgefundenen Muster zur Extraktion weiterer Relationsinstanzen und u. U. Verfeinerung des Musters (Wiederholung der Schritte 1-3 bis zur Sättigung des Prozesses).

Diese Vorgehensweise beruht auf der bereits dargestellten Annahme, dass zwischen Tuplen aufgefundene Satzbestandteile als linguistische Marker semantischer Relationen fungieren. Der von Hearst beschriebene grundlegende *bootstrapping*-Algorithmus kehrt die aus dieser Annahme abgeleitete Vorgehensweise einer musterbasierten Suche gleichsam um und nutzt nicht Muster zum Auffinden von Kontexten und Termini, sondern Termini iterativ zum Auffinden von Mustern.

Innerhalb der Terminologieforschung wurde das *bootstrapping* von *knowledge pattern* erstmalig durch Condamines und Rebeyrolle (1998) beschrieben. In einer späteren Weiterentwicklung des zuvor noch weitgehend manuell durchgeführten *bootstrappings*

(Condamines/Rebeyrolle 2001: 132 ff.) bilden die Autoren zunächst Taxonomien mit Hilfe von *Hearst patterns* sowie unter Ausnutzung der sogenannten *head characteristic*. Hiermit wird der Umstand bezeichnet, dass Mehrworttermini mit demselben syntaktischen Kopf häufig in einer semantischen Beziehung zueinander stehen. Alsdann werden diese Taxonomien miteinander verknüpft, um Tuplen zu bilden.

Bootstrapping-Ansätze für die Extraktion definitorischer Aussagen werden von Walter (2010: 235 ff.) und Reiplinger et al. (2012) beschrieben. Das von Walter (ebd.: 237 ff.) implementierte *bootstrapping*-System realisiert die folgenden Komponenten:

- manuelle *seed*-Term-Auswahl und Gewinnung neuer *seeds* mittels syntaktischer und statistischer Kriterien,
- Mustersuche mit Hilfe der *seeds* auf lemmatisierten Daten,
- Verallgemeinerung von Suchmustern ebenfalls auf Grundlage syntaktischer Kriterien,
- Bewertung von erzeugten Mustern mit Hilfe ihres *supports*[41].

Bereits im ersten Durchlauf dieses Prozesses kann Walter (ebd.: 239 ff.) 46 plausible (davon 23 neue) Suchmuster generieren, die allerdings 64 unbrauchbaren, unpräzisen oder unplausiblen Mustern gegenüberstehen. Der Autor (ebd.: 244 f.) fasst die Ergebnisse seines *bootstrapping*-Experiments für die Extraktion juristischer Definitionen wie folgt zusammen:

> „Das diskutierte Experiment hat gezeigt, dass sich schon unter Verwendung eines relativ einfachen Bootstrapping-Verfahrens Suchmuster von vergleichsweise hoher Qualität erzeugen lassen. Die durch Bootstrapping gewonnenen Suchmuster sind zu den manuell erzeugten soweit komplementär, dass sich durch die Kombination beider Musterbestände signifikant bessere Ergebnisse erzielen lassen als durch die einzelnen Suchmustermengen. […] Aufgrund unserer verhältnismäßig einfachen Umsetzung des Bootstrapping-Zyklus ist mit weiterem Verbesserungspotential durch Optimierungen in den einzelnen Schritten zu rechnen."

Die Attraktivität der Methode scheint dabei insbesondere darin zu liegen, dass sie eine deutliche Verbesserung der Abdeckung verspricht.

Reiplinger et al. (2012) gehen von 13 *seed*-Termini sowie 22 *seed*-Mustern – regulären Ausdrücken über Wort-*tokens* – aus. Die Suche nach neuen *seed*-Termini mit Hilfe der Ausgangsmuster stützt sich ebenfalls auf ein statistisches Maß. Neu gewonnene Muster werden als Sequenzen von Wortarten-*tags* repräsentiert, die Verb- und Nominalphrasen-*tags* sowie Platzhalter für andere Wortarten sowie Term-Tuplen enthalten. Zur Bereinigung der gewonnenen Musterliste werden schließlich einige einfache Heuristiken angewandt. Allerdings schließen die Autoren (ebd.: 59) an das *bootstrapping* einen manuellen Bereinigungsschritt an, in dessen Ergebnis lediglich 13 neue Muster – wie dargestellt, handelt es sich dabei um Sequenzen aus Wortarten-*tags* – erzeugt werden. Die Evaluation umfasst 100 mittels *bootstrapping* extrahierte Definitionskandidaten im Vergleich zu 100 Kandidaten, die auf Grundlage eines einfachen semantisch-syntaktischen Abhängigkeitsmusters (Terminus in Subjektposition mit nachfolgender Kopula „sein" (*to be*)) extrahiert wurden. Die Definitionskandidaten wurden durch jeweils sechs Experten auf Grundlage einer fünfstufigen Skala inhaltlich bewertet. Die Autoren finden keine signifikanten Unterschiede zwischen den beiden Extraktionsmethoden in Bezug auf die Qualität der extrahierten Definitionskandidaten, stellen allerdings fest (ebd.: 63), dass „[t]he rating experiment suggests that the concept of definition quality in a specialized domain is largely subjective […]", was obige Ausführungen zur Annotation

[41] Hierbei handelt es sich um ein Evaluationsmaß, das aus dem *data mining* bekannt ist.

definitorischer Aussagen zu stützen scheint. Eine quantitative Bewertung der Ergebnisse mit Präzision, Abdeckung und f-Maß wird nicht durchgeführt.

3.4.3 Hybride Ansätze

Hybride Ansätze kombinieren einen musterbasierten (oder heuristischen) Extraktionsschritt mit ML-Methoden zur Bereinigung der Ergebnisliste. Ein Beispiel für einen solchen Ansatz ist die Arbeit von Fišer et al. (2010). Die Autorinnen trainieren zunächst eine Reihe unterschiedlicher Klassifizierer (ebd.: 2933). Als Trainingsgrundlage dient dabei Material aus der slowenischen Wikipedia, wobei jeweils erste Sätze von Wikipedia-Artikeln als definitorisch und Sätze, die zwar mit einem Terminus beginnen, aber nicht am Beginn des jeweiligen Artikels stehen, als negative Trainingsbeispiele angesehen werden. Insgesamt werden so nahezu 20.000 Trainingsinstanzen gewonnen. Als *features* werden häufig vorkommende Wortarten-*tags* und Lemmata verwendet. Um Definitionskandidaten zu erhalten, werden alsdann aus einem slowenischen Korpus Sätze nach den folgenden Bedingungen extrahiert (ebd.):

- Der Satz beginnt mit einem in dem slowenischen Wordnet SloWNet[42] (Fišer 2009) enthaltenen Terminus und enthält einen Hypo- oder Hyperonym dieses Terminus‘ oder
- der Satz enthält mindestens zwei Termini im Nominativ oder
- der Satz enthält zwei Nominalphrasen, die ein *knowledge pattern* umschließen.

Für die Erkennung der Termini (Bedingung 2) wird ein einfaches Termextraktionsverfahren implementiert, des Weiteren werden die Untersuchungsdaten mit Informationen zu Okkurrenzen von SloWNet-Einträgen annotiert, als *knowledge pattern* wird allerdings wiederum nur die slowenische Variante des Verbs „sein“ eingesetzt. Dieser Arbeitsschritt ergibt insgesamt 1044 Definitionskandidaten, die eine durchschnittliche Präzision (als Mittelwert über den einzelnen Präzisionswerten der drei Arbeitsschritte) von 29% aufweisen.

Die auf diese Weise extrahierten Definitionskandidaten werden alsdann mit Hilfe des vorab auf den Wikipedia-Daten erlernten Klassifikationsmodells den Klassen definitorisch bzw. generisch zugeordnet. Das beste Ergebnis wird mit der Klassifikation auf Grundlage der sloWNet-annotierten Daten mit einem f-Maß von 0,5 (Präzision von 63% und Abdeckung von 42%) erreicht. Dieses Ergebnis ist allerdings schwer einzuordnen, da die Autorinnen darlegen, dass bereits die Extraktion mit SloWNet allein gerade einmal 104 Definitionskandidaten erbrachte, von denen in der Evaluation lediglich 41 tatsächlich als Definitionen akzeptiert wurden. Die Autorinnen resümieren (ebd.: 2934):

> „Manual validation of the candidates revealed the true complexity of the task. While Wikipedia definitions are mostly uniformly structured and correspond with the expected [NP] is_a [NP] form, the definition candidates extracted from the corpus display a much wider range of syntactic structures. Even more difficult was the distinction between definitions and non-definitions from the content perspective. In running text, new concepts are introduced and described in a varied and ever-changing manner. Sometimes a concept is defined by specifying what it is not, other times several concepts are defined or explained within the same context.”

Es ist nicht eindeutig zu erkennen, ob allein das aristotelische Definitionsschema als Grundlage für die manuelle Evaluation der Extraktionsergebnisse diente. Die Autorinnen heben jedoch klar hervor, dass dieses Schema allein der Vielfalt der Erscheinungsformen definitorischer Aussagen in Texten

[42] http://lojze.lugos.si/darja/slownet.html, Zugriff am 21.03.2014.

nicht gerecht wird, und scheinen zu insinuieren, dass bei einer etwas freieren Evaluation bessere Ergebnisse erzielt worden wären. Fišer et al. (2010: 2935) schlussfolgern:

> „The experiment also revealed the fuzziness of the concept of definition itself, particularly when comparing encyclopaedic definitions with those found in running texts. Not only are the latter structurally more flexible, they are also register and context dependent. Thus, a plant-infecting parasite can either be defined according to its zoological taxonomy or according to its effects on the host, environment or man; in each of these cases the defining sentence may contain a different hypernym for the target term to be defined."

Walter (2010) experimentiert ebenfalls mit ML-Algorithmen, mit denen er die Ergebnisse seines musterbasierten Verfahrens für die Extraktion deutscher Definitionen aus einem Korpus von Urteilsbegründungen verbessern möchte. Walter (ebd.: 159 ff.) vergleicht in seiner Arbeit zunächst die Performanz sequenzbasierter Suchmuster mit der von Mustern, die auch syntaktische Dependenzinformationen nutzen. Die sequenzbasierten Muster nutzen dabei die Informationstypen Wortform, Wortart, Lemma und Satzposition (ebd.: 161), die sich auf eine in ihrer Reihenfolge definierte Folge von Wörtern beziehen können. Walter setzt diese Art von Abfragen mit Hilfe des relationalen Datenbank-Managementsystems MySQL[43] um (ebd.: 164), welches die Anfragen auf eine auf *token*-Basis organisierte Datenbank abbildet. Insgesamt spezifiziert der Autor (ebd.: 167) 93 sequenzbasierte Suchmuster.

Walters (ebd.: 172) syntaktische Suchmuster definieren dagegen jeweils ein Schlüsselwort (in der Regel ein syntaktisches Prädikat), dessen syntaktischen Rahmen und die Abbildung dieses Rahmens auf Definitionsbestandteile sowie ggf. Nebenbedingungen. Die Abfragen werden XML-basiert auf Grundlage der Technologien XSLT und XPath[44] umgesetzt (ebd.: 177). Insgesamt implementiert Walter (ebd.: 178) 76 dependenzbasierte Muster. Auf seinen Testdaten erhält Walter (ebd.: 187) mit dem sequenzbasierten Verfahren ein maximales f-Maß von 0,14 (7% Präzision und 89% Abdeckung) und mit dem dependenzbasierten Verfahren ein maximales f-Maß von 0,27 (19% Präzision und 45% Abdeckung). Als Ursachen für die relativ geringe Performanz nennt der Autor (ebd.: 194 ff.) u. a. Fehler bei der linguistischen Vorverarbeitung der Daten und zu eng oder zu weit formulierte Suchmuster.

Zur Verbesserung seiner Ergebnisse testet Walter (ebd.: 207 ff.) zunächst eine Reihe einfacher Verfahren wie beispielsweise die Optimierung der Menge der verwendeten Suchmuster oder die Filterung der Ergebnisse mit einer Stopwort-Liste, die allerdings keine nachhaltige Performanzsteigerung erbringen können, da zwar die Präzision erhöht werden kann, die Abdeckung aber so stark abfällt, dass im Ergebnis kein besseres f-Maß errechnet wird. Ebenfalls eine Verschlechterung des f-Maßes ergibt die Spezifikation neuer Suchmuster zur Steigerung der Abdeckung der Extraktion (ebd.: 212 ff.).

ML-Verfahren nutzt der Autor alsdann, um optimierte Treffermengen aus seinen Extraktionsergebnissen auszuwählen. Als *features* werden dabei die in Abschnitt 3.2 genannten Eigenschaften definitorischer Aussagen verwendet. Walter testet zunächst unterschiedliche Klassifikationsverfahren und erreicht ein maximales f-Maß von 0,33 (38% Präzision und 29%

[43] http://mysql.de/, Zugriff am 21.03.2014.

[44] XPath, die XML Path Language, ist eine Sprache, mit der Abfragen über den Strukturen von XML-Dokumenten formuliert werden können. Auch für die Spezifikation von XPath existiert eine offizielle Empfehlung des *World Wide Web Consortium* (W3C): http://www.w3.org/TR/xpath/, Zugriff am 21.03.2014.

Abdeckung) mit einem sogenannten NBTree-Klassifizierer[45]. Ein weiteres Experiment ist das Ranking von Extraktionsergebnissen nach deren vermuteter Qualität, wobei „bessere“, d. h. solche Treffer, die mit höherer Wahrscheinlichkeit eine valide Definition darstellen, in der Ergebnisliste weiter oben angeordnet werden. Hierfür verwendet Walter (ebd.: 229) mit der linearen Regression[46] ein statistisches Standardverfahren. Als Merkmale dienen hierbei neben den genannten linguistischen Merkmalen auch die Ergebnisse der vorhergehenden Klassifizierung mit unterschiedlichen Algorithmen sowie Präzisionsschätzungen für die extrahierenden Suchmuster (ebd.: 228 f.). Hier wird bereits allein aufgrund linguistischer Informationen eine deutliche Verbesserung der Präzision im oberen Segment der Trefferliste erreicht. Eine solche Analyse erlaubt darüber hinaus eine Untersuchung derjenigen Merkmale, die für die Klassifikation der Definitionskandidaten tatsächlich von Bedeutung sind: Abbildung 7 ist aus der Arbeit Walters (ebd.: 232) entnommen und gibt einen Überblick über diejenigen Merkmale, die sich in verschiedenen Durchläufen des Experiments für das Ranking – erkennbar an ihren durch den Regressionsalgorithmus ermittelten Regressionskoeffizienten – als relevant erwiesen. „Positive“ Merkmale sind dabei solche, die auf das Vorliegen einer Definition hindeuten, „negative“ Merkmale dagegen treten eher mit generischen Aussagen auf. Starke Indizien für das Vorliegen einer der beiden Klassen sind Werte, denen der Algorithmus einen Koeffizienten nahe |1| zuordnet, ein Koeffizient nahe 0 deutet auf einen schwachen Effekt hin.

Entsprechend der Abbildung sind Negation, Stopwörter, definite Definienda oder aber lexikalische Beziehungen zwischen Definitionsbestandteilen relevante Negativmerkmale, wohingegen die Anordnung der Definitionsbestandteile, Kontexttreffer oder morphologische Beziehungen zwischen Definitionsbestandteilen als positive Merkmale gewertet werden können[47]. Walter (ebd.: 229 ff.) betont, dass eine abschließende Bewertung der verschiedenen Settings aufgrund seines Experiments nicht möglich sei, weist aber darauf hin, dass letztlich nur ein geringer Teil seiner *features* tatsächlich in die Regression einfloss, so dass hier sicher noch Raum für eingehendere Untersuchungen ist. Allerdings scheinen seine Ergebnisse die bisherigen Ausführungen auf Grundlage der Literatur zu relevanten und eher zweifelhaften *features* zu stützen. Die Position eines Definitionskandidaten im betrachteten Dokument spielt demnach für das Ranking kaum eine Rolle und die Definitheit des Definiendums – erinnert sei an Pearsons (1998) Postulat von Regularitäten beim Artikelgebrauch – erscheint in Walters Analyse sogar als ausgesprochenes Negativmerkmal. Nicht zuletzt erreicht Walter (2010: 231 ff.) bei einem weiteren Experiment auf deutlich größerer Datenbasis mit einem Ranking aufgrund von Präzisionsschätzungen für Suchmuster und linguistischen Merkmalen das beste Ergebnis.

Westerhout (2010: 124) wendet ein Klassifikationsverfahren auf die Ergebnisse eines musterbasierten Extraktionsschrittes für niederländische Definitionen an. Die Autorin arbeitet mit einem etwa 400.000 Wörter umfassenden Korpus, das 603 Definitionen enthält (ebd.: 95). Die Daten liegen in einem XML-Format vor und sind morpho-syntaktisch sowie mit Wortarten- und Lemma-

[45] Ein solcher Klassifizierer lernt Entscheidungsbäume, an deren Blättern zusätzlich probabilistische *Naive-Bayes*-Klassifizierer sitzen (vgl. Witten et al. 2011[3]: 456).

[46] Bei linearer Regression wird eine lineare Funktion gelernt, mit deren Hilfe numerische Werte vorhergesagt werden, indem erlernte *feature*-Gewichte linear miteinander kombiniert werden (vgl. Sachs/Hedderich 2009[13]: 650 ff.).

[47] Für eine Diskussion der Unterschiede zwischen lexikalischen und morphologischen Beziehungen vgl. Walter (2010: 231).

Gewichtung	*Konstellation (1)*	*Konstellation (2)*	*Konstellation (5)*
positiv			
↑	Definiendum vor Definiens	Präzisionsschätzung	Präzisionsschätzung
	Anaphorisches Definiendum	morph. Ähnlichkeit der Bestandteile	morph. Ähnlichkeit der Bestandteile
	definiter Genusbegriff	Definiendum vor Definiens	Definiendum vor Definiens
	Treffer im Kontext	definiter Genusbegriff	Anaphorisches Definiendum
	morphol. Ähnlichkeit der Bestandteile	Anaphorisches Definiendum	lexikalische Klassifikation
	Boostwort	Treffer im Kontext	definiter Genusbegriff
	Modifikator	TF / IDF des Definiendum	Treffer im Kontext
	TF / IDF des Definiendum		Boostwort
0			
↓	relative Dokumentposition	relative Dokumentposition	Definition kein Hauptsatz
	Definition kein Hauptsatz	Definition kein Hauptsatz	Definites Definiendum
	Subsumtionssignal	Subsumtionssignal	Subsumtionssignal
	lexikalische Ähnlichkeit der Bestandteile	Definites Definiendum	Negation
	Definites Definiendum	Negation	lexikalische Ähnlichkeit der Bestandteile
	Stoppwort (automatische Erweiterung)	lexikalische Ähnlichkeit der Bestandteile	Stoppwort (automatische Erweiterung)
	Negation	Stoppwort (automatische Erweiterung)	
negativ			

Abbildung 7: Gewichtung der Ranking-Merkmale bei linearer Regression in verschiedenen Settings nach Walter (2010: 232)

Information annotiert (ebd.: 74). Auf Grundlage dieses Formats entwickelt Westerhout (ebd.: 79) zunächst eine komplexe Extraktionsgrammatik, die Wortarteninformationen ebenso nutzt wie *knowledge patterns*, diese Informationen mit Hilfe einfacher regulärer Ausdrücke abfragt und später zu komplexeren Regeln zusammensetzt. Als Definiendum werden dabei nur satzinitiale Nominalphrasen akzeptiert (ebd.: 86). Verallgemeinernde Adverbien wie *gewoon* und *eigenlijk* (ebd.: 87) und definitionstypische Strukturen wie beispielsweise Appositionen (ebd.: 86 f.) werden beachtet. Die Extraktionsregeln sind als *rewrite*-Regeln implementiert, die dem Ursprungstext entsprechende XML-Tags hinzufügen, wenn ein Satz aufgrund der Regeln als Definition eingestuft wird (ebd.: 91 ff.).

Entsprechend ihrer Typologie definitorischer Aussagen – der in vorliegender Arbeit ausdrücklich nicht gefolgt werden soll – ergab dieser Schritt für unterschiedliche Definitionstypen die folgenden Präzisions- und Abdeckungswerte auf dem Testkorpus, das 130 Definitionen enthält (ebd.: 95 ff.):

- Definitionen mit der Kopula „sein“ (*zijn*): 34% Präzision, 79% Abdeckung, f-Maß von 0,48.
- Definitionen mit anderen Verben, die Definiendum und Definiens verbinden: 37% Präzision, 69% Abdeckung, f-Maß von 0,48.

- Definitionen, in denen Definiendum und Definiens durch Interpunktion miteinander verbunden werden: Hier enthält das Testkorpus lediglich sieben Fälle, so dass diese Daten als nachrangig erachtet werden und in der folgenden Diskussion außer Acht gelassen werden sollen.
- Definitionen, in denen Definiendum und Definiens durch Pronomen miteinander verbunden werden: Hier enthält das Testkorpus lediglich 20 Fälle, so dass diese Daten ebenfalls als nachrangig erachtet und außer Acht gelassen werden.

Da das Ziel dieses Extraktionsschritts das Erreichen einer möglichst großen Abdeckung war (ebd.: 103), wendet Westerhout alsdann ein ML-Verfahren an, mit dem sie die Präzision ihrer Ergebnisse steigern möchte. Die von ihr für ML-Experimente verwendeten *features* sind dabei die folgenden (ebd.: 127):

- *ngrams*, d. h. Sequenzen von interessierenden Elementen, hier konkret Wortarten-*tags* und morphologische Informationen,
- Informationen zum *knowledge patterns*, d. h. die Art des Musters selbst sowie Elemente aus dessen unmittelbarem Kontext (Wortarten-*tag* und morphosyntaktische Information der jeweils rechts und links benachbarten Elemente),
- Informationen zu Definiendum und Definiens, insbesondere Informationen zu bestimmten Wortarten und Phrasentypen in diesen Definitionsbestandteilen (bestimmte und unbestimmte Artikel, Eigennamen, Adjektive, Relativsätze),
- Positionsinformationen zu einem Definitionskandidaten (Position in einem Absatz, Länge dieses Absatzes, Beginn und Ende der definitorischen Satzbestandteile im Satz selbst, Häufigkeit des Definiendums vor und nach der Definition),
- Informationen zu besonderen typographischen Hervorhebungen,
- Informationen zur Wichtigkeit des Definiendums im gerade untersuchten Dokument, also Informationen über dessen terminologischen Status (ähnlich zu dem von Walter 2010 verwendeten TF-IDF-*score*).

Die Autorin führt schließlich Klassifikationsexperimente mit einem sogenannten *Balanced Random Forest Classifier* durch, der auf dem bereits eingeführten Entscheidungsbaumlernverfahren basiert (ebd.: 157):

> „The Random Forest classifier is a decision tree algorithm in which prediction is not based on one tree, but on an ensemble of trees […]. This 'forest' of trees is created by taking bootstrap samples D_i of the training data and using random feature selection for tree induction. A bootstrap sample is a sample in which the sampling instances have been selected from a data set D with replacement. This makes it possible that some instances will be contained more than once in each D_i. Predictions are made on the basis of a majority vote, that is, the class predicted by most of the trees is selected."

Ein *balancierter Random-Forest*-Klassifizierer soll nun mittels einer speziellen Methode für die Generierung der Trainingssamples für die einzelnen Entscheidungsbäume die Schwierigkeit beheben, dass in Textkorpora Definitionen normalerweise deutlich seltener auftreten als generische Aussagen (ebd.: 158 f.), d. h. er soll im Hinblick auf den Anteil von Definitionen besser balancierte Trainingssets erstellen.

Die Hinzunahme der linguistischen Informationen zu Definiendum und Definiens (*feature* 3) vermag das f-Maß für Definitionen mit „sein" auf 0,64 (62% Präzision und 67% Abdeckung) zu steigern (ebd.: 178). Für Definitionen mit anderen verbalen Konnektoren verbessern ausschließlich die Informationen zu *knowledge patterns* (*feature* 2) das f-Maß, und zwar bis auf einen Wert von 0,57 (61% Präzision und 54% Abdeckung) (ebd.: 179). Auf allen Definitionstypen gemeinsam ergibt

feature 2 das beste Ergebnis mit einem f-Maß von 0,46 (37% Präzision und 62% Abdeckung) (ebd.: 182). Die Hinzunahme der *ngram*-Information ((*feature* 1) kann die Ergebnisse für alle Definitionstypen schließlich weiter verbessern (ebd.: 198 f.).

3.4.4 Evaluation von Extraktionsergebnissen und Performanz von Extraktionsverfahren

Die Analyse von Forschungsarbeiten zur Extraktion definitorischer Aussagen lässt Schwierigkeiten bei der Evaluation von Extraktionsergebnissen aufscheinen: Mehrere Autoren unterstreichen auch dann, wenn eine quantitative Evaluation durchgeführt wird, Schwierigkeiten bei der Differenzierung zwischen *true positives* und *false positives*. Zu einem Teil ist dies sicher darauf zurückzuführen, dass nur wenigen Arbeiten, wie dargestellt, ein intersubjektiv nachvollziehbarer Annotationsschritt zugrunde liegt. Allerdings scheint es sich bei der Evaluation von automatisch extrahierten definitorischen Aussagen generell um eine komplexe Aufgabe zu handeln, die beispielsweise Regeln für den Umgang mit Teiltreffern etc. voraussetzt. Wenn keine vorannotierten Daten vorliegen, lässt sich – zusätzlich zu den Schwierigkeiten bei der Bestimmung legitimer Treffer – insbesondere die Abdeckung des Verfahrens nicht berechnen. Hierfür sind z. T. alternative Evaluationsmethoden entwickelt worden: Agichtein und Gravano (2000) versuchen dieses Problem durch eine „ideale Tabelle“ zu lösen, mit der Extraktionsergebnisse verglichen werden sollen. Malaisé et al. (2005: 39) wählen für die Berechnung der Abdeckung einen Teilabschnitt des Korpus‘ aus und vergleichen das Ergebnis der manuellen Analyse dieses Abschnitts mit den Extraktionsergebnissen. Auger und Barrière (2008: 9 f.) bieten einen Überblick über Evaluationsmethoden, in dem unterschiedliche Methoden genannt werden:

- Vergleich mit einer eigens erstellten „idealen Tabelle“ ähnlich wie bei Agichtein/Gravano (2000),
- zufällige Auswahl eines Korpusabschnitts und manuelle Auswertung,
- vollständige manuelle Auswertung der Daten,
- Vergleich mit einer bereits bestehenden Ontologie als „Gold-Standard“,
- Nutzerbefragung.

Nicht vergessen werden sollte des Weiteren, dass mehrere Autoren neben quantitativen auch qualitative Evaluationsmethoden beschreiben. In zwei Arbeiten wird zudem nicht auf Definitions-, sondern auf *token*-Ebene evaluiert.

Während das Gros der Autoren die bereits eingeführten quantitativen Evaluationsmaße Präzision, Abdeckung und f-Maß verwendet, nennt Walter (2010: 227) für die Bewertung von Ranking-Verfahren auch das Maß *mean average precision* (MAP):

11) $$MAP = \frac{\sum_{r=1}^{n} p(r) * tp(r)}{n}$$

Der Wert *n* bezeichnet dabei die Anzahl aller legitimen Treffer in der zu rankenden Ergebnismenge, *p(r)* bezieht sich dann auf die Präzision des gerankten Segments bis zu einem konkreten Treffer, während *tp(r)* eine Funktionen darstellt, die bei jedem legitimen Treffer den Wert 1 ausgibt, andernfalls den Wert 0. Neben dem Abtragen von Präzisionswerten für bestimmte Abdeckungsschwellen und deren Zusammenfassen in Präzisions-Abdeckungs-Kurven steht damit eine Evaluationsmethode zur Verfügung, mit deren Hilfe Ranking-Algorithmen anhand eines numerischen Performanzwerts miteinander verglichen werden können.

Verallgemeinernde Aussagen zur Performanz unterschiedlicher Algorithmen können dennoch nur sehr eingeschränkt getroffen werden, da in den vorgestellten Forschungsarbeiten jeweils unterschiedliche Methodologien zur Bestimmung legitimer Treffer, unterschiedliche Korpora und unterschiedliche Evaluationsmethoden zur Anwendung kommen: Ergebnisse auf Wikipedia-Daten, für die, wie dargestellt, eine gewisse, die Extraktionsaufgabe erleichternde Standardisierung der linguistischen Gestalt angenommen werden kann, können kaum mit Extraktionsexperimenten auf Grundlage von Web-Korpora verglichen werden. Experimente auf künstlich zusammengestellten Datensätzen mit relativ großen Anteilen definitorischer Aussagen – auch dies wurde bereits angemerkt – können nicht mit Experimenten auf natürlich vorkommenden Texten verglichen werden. Die von Walter (2010) untersuchten juristischen Definitionen scheinen zudem ein völlig eigenes Sonderproblem darzustellen. Nicht zuletzt erlaubt die Verwendung von *accuracy*-Werten, wie diskutiert, keine umfassende Bewertung der Güte eines Klassifikationsverfahrens.
Stark verallgemeinernd kann immerhin festgehalten werden, dass sich das Gros der dargestellten Extraktionsergebnisse in einem mittleren Bereich bewegt, womit die in einigen Arbeiten recht pauschal behauptete Überlegenheit von ML-Ansätzen – für die immerhin ein beeindruckendes Spektrum an methodischen Variationen entwickelt wurde – im Großen und Ganzen nicht bestätigt werden kann, zumal gerade Arbeiten in diesem Bereich sich häufig durch stark vereinfachende Annahmen zur linguistischen Gestalt definitorischer Aussagen auszeichnen. Generell scheint in Bezug auf die Extraktion definitorischer Aussagen aus Textkorpora das Erreichen höherer Performanzwerte, d. h. solcher Werte, die eine direkte praktische Anwendung erlauben, außerordentlich schwierig.

3.5 Zusammenfassung

Im Rahmen des in diesem Kapitel geleisteten Literaturüberblicks wurden Forschungsarbeiten aus der Terminologie sowie der Computerlinguistik und der Informationsextraktion zusammengetragen, um Antworten auf die folgenden Fragen zu finden:

- Wie wird fachspezifisches Wissen in Texten dargestellt, d.h. welche linguistischen Merkmale weisen wissenshaltige Kontexte auf?
- Wie erkennen Übersetzer fachspezifisches Wissen in Texten, d.h. gibt es überhaupt einen Konsens in Bezug auf die Relevanz fachspezifischer Aussagen in Texten für die übersetzungsorientierte Terminologiearbeit?
- Wie kann relevantes semantisches Wissen in natürlichsprachigen Texten mit Hilfe von Computerprogrammen erkannt und automatisch aus ihnen extrahiert werden?

In Bezug auf Fragestellung Nr. 1 (linguistische Merkmale wissenshaltiger Kontexte) ergibt sich ein gemischtes Bild. Als am besten untersucht kann in diesem Zusammenhang die Frage nach *knowledge patterns* gelten, mit deren Hilfe wissenshaltige Kontexte in Textkorpora erkannt werden können: Obwohl solche Suchmuster keinesfalls als unproblematisch angesehen werden können und in keiner der referierten Studien eine wirklich zufriedenstellende Performanz erbringen, wurden entsprechende Muster bereits für eine Vielzahl von Sprachen entwickelt, wobei interlinguale Ähnlichkeiten nicht von der Hand zu weisen sind. Aus dieser Feststellung ergibt sich die Hypothese, dass eine theoriegeleitete Vorgehensweise die Formulierung russischer Extraktionsmuster sinnvoll unterstützen kann.
Allerdings zeigt bereits die Betrachtung der in der Literatur diskutierten *knowledge patterns* recht deutlich, dass sich die linguistische Untersuchung wissenshaltiger Kontexte – wie bereits mehrfach angemerkt – auf einem recht oberflächlichen, zumeist rein lexikalischen Niveau bewegt. Nur wenige

Autoren verwenden syntaktische, d. h. dependenz- oder teilsatzbezogene Informationen für die Formulierung von Suchmustern und das Gros der wissenshaltigen Kontexten bzw. definitorischen Aussagen zugeschriebenen linguistischen Eigenschaften ist rein lexikalischer Natur, darüber hinaus scheint ein relevanter Anteil dieser Zuschreibungen auf bestenfalls impressionistisch zu nennenden Studien zu beruhen. Die Frage nach linguistischen Eigenschaften wissenshaltiger Kontexte bleibt also vorerst offen, gleichwohl kann die hier geleistete Diskussion bereits einige Anhaltspunkte geben (z. B. auf lexikalischer Ebene: Fehlen von Negation, Kohäsionsmitteln und Referenz, eher generische, weniger subjektive Lexik, Präpositionen und Interpunktion als Musterbestandteile; auf struktureller und grammatischer Ebene: Hauptsatztendenz, Passiv, Cluster-Bildung). Eine eingehendere Untersuchung dieses Aspekts wird in Kapitel 5 dargestellt, auch in Kapitel 6 können zu diesem Aspekt weitere Ergebnisse geliefert werden.

Für die Beantwortung von Frage 2 konnten leider keine Studien gefunden werden. Allerdings zeigen die diskutierten Annotationsstudien Probleme auf, die bei der im nächsten Kapitel zu diskutierenden Erstellung des Goldstandards zu berücksichtigen sind. Neben formal methodischen (z. B. Datenauswahl und –repräsentation) sind hier insbesondere die folgenden Aspekte zu beachten:

- Erkenntnisinteresse bzw. wissenschaftliche Zielstellung, die durch die Annotation bearbeitet werden soll (Annotationsszenario),
- Auswahl und spezifische Eigenschaften der Annotatoren,
- Umfang und Gestaltung der Annotation (z. B. Zyklen, Annotationsphasen),
- inhaltliche Aufbereitung der Annotationsrichtlinie (Informationstypen und Vorgaben),
- Treffen von Entscheidungen bei Zweifelsfällen (z. B. durch Mehrheitsvotum, Experten etc.),
- Bewertung der Annotationsqualität.

In Bezug auf Frage 3 kann zunächst auf das beeindruckend breite Spektrum an Extraktionsmethodologien und untersuchten Sprachen hingewiesen werden. Wie bereits festgestellt, scheinen die referierten Studienergebnisse die mitunter behauptete Überlegenheit von ML-Methoden gegenüber musterbasierten Ansätzen nicht vollumfänglich zu bestätigen, zumal sich viele der ML-Verfahren in der einen oder anderen Weise musterbezogener Informationen bedienen oder aber eben die Generierung von Extraktionsmustern anstreben. Angesichts der häufig in der Literatur anzutreffenden vereinfachenden Annahmen zur linguistischen Ausprägung wissenshaltiger Kontexte stellt sich die Frage, inwieweit empirisch erhobene und theoretisch fundierte linguistische Informationen Extraktionsergebnisse verbessern können. Dieser Frage wird in Kapitel 6 nachgegangen. Insbesondere wird dabei auch auf die Eignung dependenzbasierter Ansätze für die Extraktion wissenshaltiger Kontexte aus deutschen und russischen Korpora einzugehen sein, da syntaktische Informationen bislang nur in wenigen Arbeiten genutzt wurden. Vorab zu überlegen ist auch, ob ein termorientierter Ansatz im Sinne Meyers (2001) gewählt werden soll. Nicht zuletzt ist auch die Frage von Interesse, ob die korpusbasierte Extraktion wissenshaltiger Kontexte für slawische Sprachen (in unserem Fall für das Russische) tatsächlich besondere Schwierigkeiten aufweist, wie von Przepiórkowski et al. (2007) behauptet.

4. Vorbereitung: Korpuserstellung und Annotationsstudie

In vorliegender Arbeit sollen linguistische Erscheinungsformen deutscher und russischer wissenshaltiger Kontexte anhand relevanter Korpus-Okkurrenzen beschrieben und hinsichtlich ihres Informationsgehalts bewertet werden. Die linguistische Beschreibung soll dabei als Grundlage für spätere Extraktionsexperimente dienen. In diesem Kapitel wird die für beide Vorhaben erforderliche Datenbasis erarbeitet. Für die hier interessierenden Sprachen Deutsch und Russisch werden zunächst Entwicklungs- und Testkorpora erstellt. Durch manuelle Annotation gültiger Kontexte wird alsdann ein Goldstandard erarbeitet, der die Grundlage für die linguistische Untersuchung wissenshaltiger Kontexte sowie die Evaluation späterer Extraktionsexperimente bildet.

4.1 Korpuserstellung

4.1.1 Babouk

Diese Studie basiert auf *web*-Korpora, die mit Hilfe eines sogenannten Korpus-*crawlers*, und zwar mit dem *crawler* „Babouk“ (de Groc 2011), erstellt wurden. Dieser Crawler wurde im Rahmen des Forschungsprojekts TTC (*Terminology Extraction, Translation Tools and Comparable Corpora*)[48] entwickelt und verwendet ähnliche Strategien wie der *crawler* BootCat (Baroni/Bernardini 2004), allerdings mit dem wesentlichen Unterschied, dass „Babouk“ zum Zeitpunkt der Korpuserstellung standardmäßig auch für die Erstellung russischsprachiger Korpora eingesetzt werden konnte.
Auf Grundlage von durch den Nutzer vorgegebenen *seed*-URLs oder *seed*-Termini sendet „Babouk“ Suchanfragen an eine Standard-Suchmaschine (für das Russische zum Zeitpunkt der Korpuserstellung Bing, später Yandex) und durchsucht auf diese Weise das Internet nach relevanten Dokumenten. Mittels der in den Dokumenten aufgefundenen Termini wird ein fachspezifisches Lexikon erstellt bzw. das durch den Nutzer bereitgestellte Lexikon von *seed*-Termini erweitert. Die thematische Relevanz einer Internetseite wird alsdann mit Hilfe eines lexikonbasierten thematischen Filters bestimmt. Auf diese Weise findet „Babouk“ in erster Linie Webseiten, die sich durch eine hohe Terminusdichte auszeichnen.

Für die Erstellung der hier verwendeten Korpora wurde eine erste Menge an Texten mittels einer relativ kleinen Anzahl initialer *seed*-Termini gecrawlt. Aus diesen wurden manuell in einem iterativen Prozess neue *seed*-Termini für weitere *crawls* extrahiert. Das deutsche Entwicklungskorpus wurde ebenfalls im Rahmen des TTC-Projekts erstellt und der Autorin zur Verfügung gestellt.

4.1.2 Lexikographische Korpora und Web-Korpora

Bei der Nutzung von Korpora bzw. *web*-Korpora für terminologische Belange sind Fragestellungen zu beachten, die mit der Natur der verwendeten Korpora zusammenhängen. Während für die praktische Terminographie dabei die Frage nach der Qualität und Zuverlässigkeit der aus einem

[48] http://ttc-project.eu/, Zugriff am 28.03.2014.

Korpus extrahierten Information im Vordergrund steht, muss für wissenschaftliche Studien der Grad der Verallgemeinerbarkeit der durchgeführten Experimente sowie der dabei ermittelten Ergebnisse klar nachvollziehbar sein. In der klassischen Lexikographie, die auf das Erarbeiten allgemeingültiger Beschreibungen lexikalischer Einheiten abzielt, werden ähnliche Fragestellungen in der Regel durch eine streng regulierte Korpuserstellung und ein genaues Monitoring der Korpusinhalte beantwortet. Demgegenüber betont der *web-as-corpus*-Ansatz die schiere Datenmenge großer *web*-Korpora, die mittel automatischer Verfahren erstellt („gecrawlt") wurden. Die sich aus den beiden unterschiedlichen Ansätzen ergebenden theoretischen und methodischen Implikationen sollen hier mit Blick auf das in dieser Arbeit verfolgte Forschungsvorhaben kurz diskutiert werden.
Traditionell werden Textkorpora als elektronische Textsammlungen definiert, „[...] built according to explicit design criteria for a specific purpose [...]" (Atkins et al. 1992: 1). Dieser eher lexikographischen Tradition entsprechend definiert auch Sinclair (1991: 62) ein Textkorpus als „[...] collection of naturally occurring language text, chosen to characterize a state or variety of a language" – in einem solchen Untersuchungsansatz ist die in dem Korpus dokumentierte Sprache selbst das Untersuchungsobjekt und nicht lediglich Mittel zu einem anderen Zweck (wie beispielsweise der Extraktion wissenshaltiger Kontexte). Ziel der Korpuserstellung in solchen Ansätzen ist das Generieren einer Datenpopulation, die für eine bestimmte – dem Forschungszweck entsprechende – Grundgesamtheit *repräsentiv* ist oder diese zumindest frei von künstlichen Verzerrungen repräsentieren kann. Die Überprüfung existierender Korpora auf ihre Repräsentativität bzw. Balanciertheit erweist sich dabei jedoch als theoretisch und methodisch nur schwer zu lösendes Problem (Atkins et al. 1992: 6, Hervorhebungen im Original):

> „There is much talk of a 'balanced corpus' as a *sine qua non* of corpus analysis work: by 'balanced corpus' is meant (apparently) a corpus so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study. At present, corpus 'balance' relies heavily on intuition […]. In our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus – however 'unbalanced' – to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts. It would be shortsighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' simply because the corpus used cannot be proved to be 'balanced'."

Atkins et al. (ebd.: 7 ff.) nennen einen Kriterienkatalog, der für die Erstellung und Beschreibung der in Textkorpora vorkommenden Dokumente verwendet werden kann. Auch Engwall (1994) führt ähnliche Kriterien an. Tabelle 5 gibt einen Überblick über einige ausgewählte Beschreibungskriterien zur Charakterisierung der Dokumente in einem Textkorpus:

Attribut	**Beschreibung**
Modalität	Existenzform des Textes, z. B. mündlich, schriftlich etc.
Herkunft	Urheber des Textes, z. B. Einzelperson, Gruppe etc.
Textqualität (*preparedness*)	Arbeitsaufwand bei der Texterstellung, z. B. Spontantext, freie Rede nach Notizen etc.
Stil	z. B. Prosa, Vers etc.
Textsorte	z. B. Roman, Vorlesung etc.
Faktizität	an reale Tatsachen gebundene versus fiktive Texte
Kommunikationssituation	z. B. Ausbildung, Arbeitsplatz, Freizeit etc.
Funktion	z. B. narrativ, informativ, anleitend etc.
Thema	Textgegenstand

Spezialisierungsgrad	allgemeine Texte versus Fachtexte
Veröffentlichungszeitpunkt	Alter (Aktualität) des Textes

Tabelle 5: Überblick über ausgewählte Kriterien zur Beschreibung von Textkorpora nach Atkins et al. (1992)

Im Falle von *web*-Korpora ist eine derart detaillierte Beschreibung des Korpusinhalts insbesondere mit zunehmender Größe kaum noch zu leisten und erscheint je nach Disziplin auch nicht immer notwendig. Dies kann sich jedoch dahingehend als problematisch erweisen, als Online-Texte in der Regel eine geringere Textqualität erwarten lassen und entsprechende Phänomene dokumentiert werden sollten. So verweist Fletcher (2004: 192) darauf, dass online – zusätzlich zum Problem der oft nicht feststellbaren Autorenschaft – bestimmte Textsorten, etwa journalistische und akademische sowie kommerzielle Texte, überwiegen, so dass bestimmte Repräsentationserfordernisse durch Online-Texte u. U. nicht zu erfüllen seien. Internettexte seien zudem häufig von vergleichsweise schlechter Qualität (ebd.):

> „Overall, lower standards of form and content verification prevail than in printed sources. Web pages often contain a significant amount of 'noise' (i.e. language which is fragmentary, repetitive, formulaic, or ill-formed, and often entire documents which have no cohesive text)."

Für die Arbeit mit *web*-Korpora schlägt Fletcher daher eine Reihe z. T. elaborierter Filterungs- und Bereinigungsverfahren vor.

Kilgarriff und Grefenstette (2003: 333 f.) merken im Widerspruch zu klassisch-lexikographisch orientierten Ansätzen der Korpuslinguistik und unter Bezugnahme auf Manning und Schütze (1999) an, dass gerade in der statistischen Sprachverarbeitung die Balanciertheit und genaue Dokumentation eines Korpus' nur eine untergeordnete Rolle spielten: Für statistische Ansätze sei insbesondere die Menge an Inputdaten ausschlaggebend, nicht deren interne Zusammensetzung, und die Frage nach dem Status eines Korpus' sei von der Frage nach der Qualität des entsprechenden Korpus' zu trennen. Dieser pragmatische Ansatz mündet in der Definition eines Textkorpus' als „[...] a collection of texts when considered as an object of language or literary study" (ebd.: 334). Die Frage nach der Korrektheit der aus Korpora gewonnenen Informationen beantworten die Autoren (ebd.: 342) mit dem Verweis auf statistisch signifikante Distributionsunterschiede, auf die bei der Arbeit mit *web*-Korpora zu rekurrieren sei: „All the 'erroneous' forms exist, but much less often than the 'correct' forms. The Web is a dirty corpus, but expected usage is much more frequent than what might be considered noise". Eine solche Sichtweise relativiert Repräsentativitäts- und Balanciertheitsfragen und favorisiert den Umgang mit größeren und damit auch unterdefinierten Datenmengen. Ziel entsprechend ausgerichteter Untersuchungen scheint dabei nicht mehr die Beschreibung eines Ausschnitts einer vorab definierten Sprachwirklichkeit zu sein, vielmehr wird die online vorfindbare Sprache als neuer Datentyp betrachtet, aus dem – unter Beachtung der vorhandenen Eigenheiten und Strukturen der Daten selbst – die gewünschten Informationen zu extrahieren sind. Ein *web-as-corpus*-Ansatz entzieht damit die Struktur des Korpus' zumindest in Teilen der Kontrolle des Korpuserstellers und überträgt ihm die Aufgabe, Methoden zu entwickeln, welche die zielgerichtete Arbeit mit den online vorfindbaren Daten für ein üblicherweise über die reine Deskription sprachlicher Eigenschaften hinausgehendes Forschungsvorhaben erlauben (ebd.: 343):

> „The Web is not representative of anything else. But nor are other corpora, in any well-understood sense. Picking away at the question merely exposes how primitive our understanding of the topic is

> and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it might be modeled."

Die korpusgestützte Terminographie scheint in dieser Hinsicht einen Mittelweg einzuschlagen, indem sie weniger die Erstellung balancierter und wohldokumentierter Korpora als vielmehr die produktive Kritik der für Experimente verwendeten Ressourcen und Methoden in den Mittelpunkt stellt. So betonen etwa Ahmad und Rogers (2001: 726 f.) im Hinblick auf die Terminologieextraktion:

> „The evaluation of any term-extraction method must go hand in hand with an evaluation of the corpus that is being analyzed, not simply as in traditional terminology work reference […] to the authority of the authors of texts, but also with regard to the structure and processing of the corpus."

Eine solche Vorgehensweise trägt einerseits der Tatsache Rechnung, dass terminologisch relevante Informationen aus einer Vielzahl unterschiedlichster Quellen zusammengetragen werden können und die Balanciertheit eines Textkorpus' für ein solches Vorhaben weniger ausschlaggebend erscheint als etwa für soziolinguistische oder sprachdiachrone Untersuchungen. Wie bereits einleitend festgestellt, wird das Internet von Terminologen, Übersetzern und Dolmetschern seit Jahren als Informationsquelle genutzt und eine korpuslinguistisch ausgerichtete Terminographie sollte diesem Umstand Rechnung tragen. Andererseits liegt es auf der Hand, dass eine wesentliche Voraussetzung für eine fruchtbare Korpuskritik eine umfangreiche Dokumentation des fraglichen Korpus' ist.

In diesem Zusammenhang spielt auch der Umstand eine Rolle, dass das hier interessierende Untersuchungsobjekt – wissenshaltige Kontexte – voraussichtlich nicht gleichmäßig über alle in einem Korpus vorkommenden Textsorten verteilt ist und, sofern die Beobachtungen Fletchers (2004) zutreffen, nur in einem geringen Teil der im Internet auffindbaren Texte überhaupt vorliegt: Wissenshaltige Kontexte vermitteln Wissen im Zusammenhang konkreter Kommunikationserfordernisse. Es liegt auf der Hand, dass eine solche Wissensvermittlung nur in bestimmten, oft asymmetrischen Kommunikationssituationen stattfindet. Daher kann die Hypothese formuliert werden, dass wissenshaltige Kontexte überwiegend in Kommunikationssituationen vorzufinden sind, in denen es um einen Wissenstransfer geht, also etwa in der Kommunikation zwischen Experten und Nicht-Experten. Demgegenüber kann in Situationen der Experte-zu-Experten-Kommunikation von einem deutlich höheren Maß an Implizität und einem geringeren Maß an Explizität, also einer geringeren Dichte an Erklärungen und Definitionen und damit wissenshaltigen Kontexten ausgegangen werden (vgl. Meyer 2001: 299):

> „In our experience, the number of KRCs varies depending on the text type. Predictably, it appears that the higher the gap between the writer's and the reader's level of domain knowledge, the greater the number of KRCs that will be found."

Etwas differenzierter unterscheidet Pearson (1998: 36 ff.) zwischen den Kommunikationssituationen der Experten-Experten-, Experten-Anfänger-, der Kenner-Laie- (*relative expert to the uninitiated*) sowie der Lehrer-Schüler-Kommunikation. Spezifische Termini einschließlich ausführlicher Erläuterungen seien lediglich in den *settings* der Experten-Anfänger- und Lehrer-Schüler-Kommunikation zu erwarten. Typische Vermittlungsformen seien Lehrbuchtexte. Die detaillierte Dokumentation des Vorkommens wissenshaltiger Texte in Korpora ermöglicht die empirische Überprüfung einer solchen Hypothese. Im Zusammenhang mit der hier zu beschreibenden Erstellung der Datenbasis ist indes zunächst zu dokumentieren, welche Arten von Texten in den mit „Babouk"

gecrawlten Korpora sowie dem auf deren Grundlage erstellten Goldstandard vorkommen. Im Sinne von Ahmad und Rogers (2001) soll die Dokumentation der verwendeten Korpora in vorliegender Arbeit damit weniger der Überprüfung der Balanciertheit und Repräsentativität der Korpora dienen – ein, wie aus vorstehenden Bemerkungen ersichtlich, nicht nur aufgrund der beschriebenen theoretischen Unschärfen, sondern auch in Anbetracht des wahrscheinlich ungleichmäßigen Auftretens wissenshaltiger Kontexte problematisches Unterfangen –, sondern den gezielten Test der aufgestellten Hypothesen sowie eine empirisch fundierte Korpus- und Methodenkritik ermöglichen.

4.1.3 Dokumentation der erstellten Korpora

Mit Hilfe von „Babouk“ wurden vier Textkorpora erstellt. Jedes Dokument in diesen Korpora ist mit einem XML-Header versehen, der Metadaten enthält. Tabelle 6 gibt einen Überblick über einige Charakteristika der erstellten Korpora. Dabei handelt es sich – mit Ausnahme des russischen Entwicklungskorpus‘, bei dessen Erstellung aufgrund der insgesamt recht geringen Textmenge domänenfremde und doppelt vorkommende Texte noch manuell entfernt werden konnten, um den Anteil relevanter Texte zu erhöhen, – um weitgehend unbereinigte Korpora, die insbesondere irrelevantes Material enthalten können. Als Top-5-Webseiten sind in Tabelle 6 diejenigen Internetseiten ausgewiesen, von denen jeweils die meisten Texte von „Babouk“ in das Korpus übernommen wurden:

Korpus	**Anzahl der Wörter im Korpus**[49]	**Onlinequellen**	**Fachgebiete**
Entwicklungskorpus Russisch (EKR)	353.094	139 Texte von 51 verschiedenen Webportalen, *Top 5*: • roman.by • niva-chevy.narod.ru • autoprospect.ru • 2101vaz.ru • hyundai-sto.ru	(weitgehend) monothematisch (KFZ)
Testkorpus Russisch (TKR)	1.099.309	421 Texte von 124 verschiedenen Webportalen, *Top 5*: • dic.academic.ru • roman.by • ru.wikipedia.org • www.pereplet.ru • wsyachina.narod.ru	Verschiedene Themen-bereiche[50]: • **Atomenergie: 44 Texte** • KFZ: 42 Texte • Physik und Chemie: 37 Texte • Energieversorgung: 31 Texte • Ingenieurwesen: 29 Texte • Umwelt: 27 Texte • alternative Energie: 22 Texte • IT: 13 Texte

[49] Die in dieser Spalte angegebenen Werte wurden mit Hilfe der wc-(word count)-Funktion in Linux auf Grundlage der nicht bearbeiteten Korpora ermittelt und beziehen sich entsprechend auf Zeichenketten (einschließlich nicht-alphabetischer Zeichen), die durch Leerzeichen begrenzt werden.

[50] Die Angabe bezieht sich auf ein Korpussample von 260 Texten.

			• Militärtechnik: 10 Texte • Raumfahrt: 5 Texte
Entwicklungskorpus Deutsch[51] (EKD)	993.214	757 Texte von 381 verschiedenen Webportalen, *Top 5*: • www.wind-energie.de • www.kaufen.com • cgi.ebay.de • www.oem-ag.at • de.wikipedia.org	(weitgehend) monothematisch (Windenergie)
Testkorpus Deutsch (TKD)	7.268.715	2270 Texte von 174 verschiedenen Webportalen, *Top 5*: • de.wikipedia.org • www.tutorials.de • de.domotica.net • www.chip.de • www.solarserver.de	Verschiedene Themenbereiche[52]: • **IT: 62 Texte** • Alternative Energie: 57 Texte • Energieversorgung: 30 Texte • Ingenieurwesen: 22 Texte • Physik und Chemie: 21 Texte • KFZ: 6 Texte • Design und Fotografie: 6 Texte • Atomenergie: 5 Texte • Recht: 3 Texte • Umwelt: 3 Texte • Geographie: 2 Texte • Militärtechnik: 1 Text • Biologie: 1 Text • Politik: 1 Text

Tabelle 6: Überblick über die erstellten russischen und deutschen Textkorpora

Die Unterschiede zwischen den hier dokumentierten Korpora sind beträchtlich: Wurden bei Erstellung des EKR noch domänenfremde Texte manuell aus dem Korpus entfernt (was insbesondere die geringe Größe des Korpus‘ erklärt), wurde auf solche Maßnahmen bei Erstellung der weiteren Korpora in Anbetracht der Datenfülle verzichtet. Des Weiteren fällt in Bezug auf alle vier Korpora auf, dass „Babouk“ eine große Menge an Texten von Sammelportalen wie *roman.by* (einer weißrussischen Seite, auf der Referate und Hausarbeiten abgerufen werden können, vergleichbar mit *hausarbeiten.de*), *dic.academic.ru* (einem Portal für Wörterbücher und Enzyklopädien), *wsyachina.narod.ru* (einem Portal mit Artikeln zu unterschiedlichen Fachthemen), *tutorials.de* und natürlich *Wikipedia*[53] abruft sowie allgemein viel Material von Seiten mit einer tiefen Substruktur sammelt.

[51] Dieses Korpus wurde im Rahmen des TTC-Projekts gesammelt und der Autorin für ihre Forschungsarbeit zur Verfügung gestellt. Die neueste Version dieses Korpus‘ steht online unter folgendem Link zum Download bereit: http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html, Zugriff am 28.03.2014.

[52] Die Angabe bezieht sich auf ein Korpussample von 220 Texten.

[53] Die Webseite *de.domotica.net* scheint derzeit nicht erreichbar zu sein, Zugriffsversuch am 24.03.2014.

Die hier angegebenen Daten erlauben die Formulierung von Hypothesen zur Qualität der erstellten Korpora im Hinblick auf die Aufgabenstellung dieser Arbeit. So handelt es sich bei dem russischen Entwicklungskorpus aufgrund der manuellen Selektion von Texten um ein zwar kleines, aber fachlich deutlich spezialisiertes Korpus. Das russische Testkorpus ist größer als das Entwicklungskorpus und deckt unterschiedliche Fachgebiete scheinbar recht gleichmäßig ab – gute Voraussetzungen für eine zumindest ansatzweise Prüfung der Übertragbarkeit der entwickelten Extraktionsmethode auf neue Fachgebiete.
Das deutsche Entwicklungskorpus ist zwar monothematisch und relativ umfangreich, die hohe Anzahl unterschiedlicher Internetquellen des Korpus‘ scheint aber auf eine mangelnde fachliche Spezialisierung hinzudeuten – gerade im Hinblick auf den großen Anteil rein kommerzieller Webseiten wie *ebay.de* und *kaufen.com*. Aus der im Vergleich zum russischen Testkorpus hohen Anzahl von Texten bei annähernd gleicher Wortzahl lässt sich zudem schließen, dass ein großer Anteil der enthaltenen Texte sehr kurz und damit für die Extraktion wissenshaltiger Kontexte eher ungeeignet ist, da KRCs eher in längeren Explikationszusammenhängen zu erwarten sind. Das deutsche Testkorpus schließlich ist zwar inhaltlich diversifiziert, weist aber einen hohen Anteil an Wikipedia-Texten auf.

4.1.4 Statistische Auswertung der erstellten Korpora

4.1.4.1 Vergleich der Entwicklungs- und Testkorpora mit Referenzkorpora

Zusätzlich zu einer Dokumentation der erstellten Korpora anhand der zur Korpuserstellung genutzten Quellen und der in den Korpora abgedeckten Fachgebiete ist auch die Beschreibung der Korpora anhand statistischer Messwerte möglich. Diese können insbesondere unter Bezugnahme auf Vergleichskorpora Informationen über sprachliche Eigenschaften der untersuchten Korpora liefern und damit Ansätze zu einer linguistisch-funktionalen Beschreibung und Einordnung insbesondere hinsichtlich der Frage, ob es sich um inhaltlich spezialisierte Texte mit einem gewissen Fachlichkeitsgrad handelt, derselben bieten. So nutzen etwa Teich und Fankhauser (2010) die folgenden Messwerte, um das „Darmstadt Scientific Text Corpus“[54] (DaSciTex, vgl. Holtz/Teich 2009) mit dem allgemeinsprachlichen F-LOB-Korpus[55] (Hundt et al. 1999) zu vergleichen:

- Relative Frequenz von Nomina, Verben und Adverbien in den untersuchten Korpora,
- Standardisierte *type token ratio* (STTR),
- Durchschnittliche Wortzahl pro Teilsatz.

Zur Messung der STTR wird dabei das Korpus in Teilabschnitte mit gleicher *token*-Zahl aufgesplittet und in diesen die Anzahl der verwendeten Lemmata gezählt. Aus den erhaltenen Werten wird anschließend der Durchschnitt gebildet (vgl. Bowker/Pearson 2002: 110) – STTR wird also als Maß für den Umfang des in einem Korpus enthaltenen Vokabulars verwendet und entsprechend als Indikator für die Unterscheidung zwischen unterschiedlichen Textarten herangezogen (Altmann 1988: 88). Diese Information kann wiederum – ebenso wie relative Frequenzen bestimmter Wortarten und Teilsatzlängen – zu Annahmen über stilistische Eigenschaften des untersuchten Korpus in Beziehung gesetzt werden. Im Vergleich zu F-LOB zeigt etwa DaSciTex einen deutlich niedrigeren STTR-Wert sowie eine höhere relative Frequenz von Nomina und eine geringere relative

[54] http://www.linglit.tu-darmstadt.de/index.php?id=lingpro_deliverable, abgerufen am 15.5.2014.
[55] http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html, abgerufen am 9.1.2014.

Frequenz von Adverbien, was die Autoren der Studie (Teich/Fankhauser 2010: 235) darauf zurückführen, dass DaSciTex wissenschaftliche Texte enthält:

> „If DaSciTex represents scientific writing, then the texts contained should exhibit some typical properties of this meta-register, such as (relative) abstractness, technicality and informational density [...], which are not exhibited by a registerially mixed corpus such as FLOB [sic!].“

Als informativste Maße bei der Unterscheidung zwischen dem fachsprachlichen DaSciTex und dem allgemeinsprachlichen F-LOB erweisen sich in der Studie von Teich und Fankhauser die STTR sowie die relative Frequenz von Nomina (NN) und Adverbien (ADV). Auch Walter (2010: 152) zeigt anhand eines Vergleichs eines Korpus‘ juristischer Fachtexte mit einem Korpus von Artikeln aus der Tageszeitung „Frankfurter Rundschau”, dass Fachtextkorpora eine geringere *type token ratio* aufweisen als allgemeinsprachliche Korpora.
Allerdings ist die STTR ebenso wie ihre Grundform, die *Type-Token*-Relation (TTR), welche die Anzahl von Typen (z. B. Lemmata) in einem Korpus zur Anzahl der *tokens* in Beziehung setzt, nicht unumstritten. Als problematisch gilt dabei insbesondere, dass die TTR von der Länge des untersuchten Korpus‘ abhängig ist, wohingegen im Falle der STTR die Festlegung der *sample*-Größe eine willkürliche Entscheidung ist und linguistische Intuitionen bzw. Hypothesen nicht abzubilden vermag (Wimmer 2005: 365):

> „Would it, perhaps, be more adequate to divide the text in several parts, compute the TTR for the parts separately and at last take a weighted average of an index or of the parameters of a curve? Perhaps, but the decision concerning the length of the parts would be so arbitrary, that no better result could be obtained.”

Vor diesem Hintergrund scheint es sinnvoll, statt der von Teich und Fankhauser (2010) verwendeten STTR auf die TTR zurückzugreifen, um die erstellten Korpora hinsichtlich ihrer Vokabulargröße jeweils einem Referenzkorpus gegenüberzustellen, wobei in diesem Fall die Vergleichsdaten eine identische Länge (d. h. eine identische Anzahl von *tokens*) aufweisen müssen, um das Problem der Abhängigkeit der TTR von der Länge des untersuchten Korpus‘ zu minimieren und einen direkten Vergleich der untersuchten Datensätze zu ermöglichen. Zwar erlaubt das auf diese Art und Weise ermittelte Ergebnis beispielweise für das sehr kleine EKR keine Aussage darüber, inwieweit sich dessen Vokabulargröße von der des wahrscheinlich deutlich größeren Referenzkorpus‘ insgesamt oder aber der des ebenfalls größeren TKR unterscheidet. Allerdings kann der Vergleich mit Referenzdaten, deren Charakteristiken bekannt oder messbar sind, dabei helfen, die erstellten Korpora zumindest ansatzweise hinsichtlich ihrer Fachlichkeit zu bewerten. Dabei wird mit Teich und Fankhauser (2010) angenommen, dass ein geringeres *type-token*-Verhältnis ein Indikator für ein eingeschränktes Vokabular aufgrund inhaltlicher Spezialisierung ist. Ein größerer Anteil an Nomina wird dann als Indikator für die häufigere Verwendung von Termini und Nominalisierungen und mithin ebenfalls als Hinweis auf den fachlichen Charakter der im Korpus versammelten Texte interpretiert. Ebenfalls mit Teich und Fankhauser wird schließlich ein geringerer Anteil von Adverbien als Merkmal von Fachtexten interpretiert.
Zur Charakterisierung der oben beschriebenen Korpora sollen für die russischen und deutschen Entwicklungs- und Testkorpora die genannten Maße kontrastiv zu Referenzkorpora erhoben werden.

Als russisches Referenzkorpus soll das „Russian Internet Corpus“ (RIC, Sharoff 2006) dienen[56], das mittels eines dem von „Babouk“ angewandten ähnlichen Verfahrens erstellt wurde. Sharoff (ebd.: 81 ff.) stellt dar, dass das RIC zu 40% aus Texten besteht, die sich an ein allgemeines Publikum richten, und zu 46% aus Texten, die sich an ein vorinformiertes (aber nicht über Expertenwissen verfügendes) Publikum richten. Mithin kann für das RIC von einem mäßigen Fachlichkeitsgrad ausgegangen werden. Für das Deutsche wird ein *crawl* von Nachrichtentexten des Leipziger Korpusportals (NCL, Quasthoff et al. 2006) aus dem Jahr 2011[57] verwendet. Das NCL stellt eine Sammlung randomisierter Sätze aus deutschen Tageszeitungen dar (ebd.: 1799).
Da die russischen und deutschen Referenzkorpora deutlich größer sind als die mit „Babouk“ erstellten Fachkorpora, werden für den Korpusvergleich jeweils Unterabschnitte des RIC und des NCL entsprechender Größe verwendet, nicht die kompletten Korpora. Der Empfehlung von Altmann (1988: 85) folgend werden die zu untersuchenden Texte vorher mit Hilfe von TreeTagger (Schmid 1994) lemmatisiert, da andernfalls die *type-token*-Relation für das Russische und Deutsche – Sprachen mit mittelstark bis stark ausgeprägten Flexionsparadigmen – lediglich als Maß für die Häufigkeit unterschiedlicher Wortformen, nicht aber lexikalischer Einheiten angesehen werden kann. Dennoch ist zu betonen, dass aufgrund von gerade im terminologischen Bereich häufigen Lemmatisierungsfehlern auch die TTR nur als Näherungswert angesehen werden kann. Auch für die Wortartenannotation ist mit einer gewissen Fehlerquote zu rechnen. Um die Untersuchung auf das Lexikon zu beschränken, werden zudem nicht-alphabetische Zeichen aus den Korpora entfernt.
Anhand der getaggten und lemmatisierten Datensätze wird alsdann ein Vergleich der Häufigkeiten von Wortarten sowie von Lemmata durchgeführt. Mit Hilfe statistischer Testverfahren werden die ermittelten Verteilungsunterschiede auf Signifikanz geprüft. Konkret werden die erhobenen Frequenzdaten für die beiden Datensätze als Vierfeldertafeln (vgl. Sachs/Hedderich 2009[13]: 557 ff.) modelliert. Ein Beispiel für eine solche Vierfeldertafel ist in Tabelle 7 für das Merkmal „Anzahl der Lemmatypen“ anhand der beiden russischen Datensätze dargestellt.

Merkmal	**EKR**	**RIC**	**Gesamt**
Anzahl der Lemmatypen im untersuchten Datensatz, d. h. Anzahl derjenigen Lemmata, die mindestens einmal vorkommen	19.310	26.103	45.413
Anzahl der verbleibenden Lemma-*tokens* (wiederholte Lemmata)	301.496	294.703	596.199
Lemma-*tokens* gesamt[58]	320.806	320.806	634.766

Tabelle 7: Beispiel für eine Vierfeldertafel

Als statistisches Verfahren zur Prüfung der Frequenzunterschiede auf Signifikanz kann für eine solche Vierfeldertafel mit dem Chi-Quadrat-Test (ebd.) ein statistisches Standard-Testverfahren verwendet werden. Als Nullhypothese H_0 wird dabei angenommen, dass die untersuchten Datensätze bezüglich des interessierenden Merkmals als Stichproben aus derselben Grundgesamtheit (z. B. der Grundgesamtheit der russischen Sachtexte) angesehen werden können

[56] http://corpus.leeds.ac.uk/internet.html, Abruf am 9.1.2014. Das Korpus wird mittlerweile neu aufgesetzt und steht daher nicht mehr zum Download zur Verfügung.
[57] http://corpora.informatik.uni-leipzig.de/download.html, Zugriff am 28.03.2014.
[58] Für diese Untersuchung wurden nicht-alphabetische Zeichen aus den Datensätzen entfernt.

und dementsprechend homogene Verteilungen des interessierenden Merkmals aufweisen. Alternativ (Alternativhypothese H_1) wird davon ausgegangen,
dass die Datensätze aus zwei unterschiedlichen Grundgesamtheiten (z. B. russische Fachtexte vs. russische allgemeinsprachliche Texte) stammen und sich entsprechend bezüglich des untersuchten Merkmals unterscheiden. Wird die Nullhypothese nicht verworfen, gilt die Häufigkeit einer Wortart oder eines Lemmas als stochastisch unabhängig vom untersuchten Datensatz. Wird dagegen die Alternativhypothese bestätigt, kann davon ausgegangen werden, dass die untersuchten Korpora jeweils spezifische Verteilungen dieses Merkmals aufweisen (ebd.).
Für diese Untersuchung wird die korrigierte Variante des Chi-Quadrat-Tests verwendet, die als robuster gilt als die klassische Variante (vgl. Evert 2009: 1235). Die Formel für den Chi-Quadrat-Test ist in 12) dargestellt, wobei N die Gesamtgröße der beiden untersuchten Datensätze, O_{11} den Eintrag in der ersten Zeile der ersten Spalte der Vierfeldertafel (O_{12} dagegen den Eintrag der ersten Zeile und zweiten Spalte etc.), R_1 die Summe der ersten Zeile der Tabelle und C_1 die Summe der ersten Spalte bezeichnen (ebd., R_2 und C_2 entsprechend).

12) $$\chi^2 = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - \frac{N}{2})^2}{R_1 R_2 C_1 C_2}$$

Je nach gewähltem Signifikanzniveau kann die Nullhypothese abgelehnt werden, wenn der mit Hilfe des Tests errechnete p-Wert, der ein „[…] Evidenzmaß für die Glaubwürdigkeit der [Nullhypothese] […]“ (Sachs/Hedderich 2009[13]: 378) darstellt, einen bestimmten Wert unterschreitet. Allerdings setzt der Chi-Quadrat-Test auf Grundlage einer Vierfeldertafel voraus, dass keine der erwarteten Häufigkeiten E_{ij} in der Tabelle (ein Rechenbeispiel gibt Evert 2009: 1231) den Wert 5 unterschreitet (Graf et al. 1998[3]: 196). In diesen Fällen wird anstelle des Chi-Quadrat-Tests der exakte Fisher-Test durchgeführt (Sachs/Hedderich 2009[13]: 559, 575 ff.). Da es sich hierbei um einen exakten Test handelt wird keine Prüfgröße angegeben, vielmehr wird zur Entscheidung über die mögliche Ablehnung der Nullhypothese allein der p-Wert verwendet.
Anzumerken ist in diesem Zusammenhang, dass die Verwendung statistischer Hypothesentests und insbesondere des Chi-Quadrat-Tests in der Korpuslinguistik auf Kritik gestoßen ist. Gries (2005: 278) verweist darauf, dass „[…] a large enough sample will always generate a significant result […]“, womit gemeint ist, dass bei Untersuchung großer Datenmengen, wie es in der Korpuslinguistik der Normalfall ist, auch marginale Unterschiede statistische Signifikanz erlangen können. Ein Maß für die Nicht-Marginalität, d. h. die praktische Bedeutung eines beobachteten Unterschieds zwischen zwei Stichproben, ist die sogenannte Effektstärke. Die Frage, welche Effektstärken und welche Signifikanzniveaus als Richtwerte anzusetzen sind, d. h. wann ein statistischer Unterschied nicht nur signifikant, sondern auch praktisch bedeutsam ist, wird jedoch in der Korpuslinguistik ebenso wie in anderen Disziplinen kontrovers diskutiert. Ganz allgemein kritisieren etwa Rosenthal et al. (2000: 4 f.) die konventionelle Verwendung des 5%-Signifikanzniveaus:

> „It may not be an exaggeration to say that for many Ph.D. students, for whom the .05 alpha has acquired a kind of mystique, it can mean joy, doctoral degrees, and tenure track positions at major universities if their dissertation *p*s are less than .05.”

Im selben Band (ebd.: 25 ff.) diskutieren die Autoren die Interpretation von Effektstärken und fordern eine sachgerechte, d. h. an praktischen Erfordernissen orientierte Abwägung bei der Festlegung einer minimalen Effektstärke. Sachs und Hedderich (Sachs/Hedderich 2009[13]: 379)

empfehlen für große Stichproben die Verwendung eines 0,1%-Signifikanzniveaus. Gries (2005: 280) verwendet für die Beschreibung der Effektstärke eine Reihe von Indizes, beispielsweise den Cramer-Index φ. Als weiteres Maß, mit dem ein statistischer Unterschied quantifiziert werden kann, nennt Gries (2009: 174) das Chancen-Verhältnis (*odds ratio*). Baroni und Evert (2009: 793) charakterisieren die minimale Größe des Unterschieds zwischen Stichproben mit Hilfe eines Konfidenzintervalls und beschreiben ein Beispiel, in dem der minimale Unterschied zwischen zwei Stichproben, d.h. die untere Grenze des entsprechenden Konfidenzintervalls, 3% beträgt, was aus statistischer Sicht als gering einzuschätzen ist, in einem korpuslinguistischen *setting* jedoch ohne Weiteres Frequenzunterschiede in einem dreistelligen Bereich beschreiben kann (vgl. wiederum die Diskussion in Rosenthal et al. 2000: 25 ff.). Tatsächlich scheint es keine Richtwerte für Effektstärken in der Korpuslinguistik zu geben: Werden selten auftretende linguistische Phänomene untersucht, bleibt fraglich, inwiefern die ermittelten Effektstärken mit jenen für hochfrequente Phänomene verglichen werden können.

Angesichts der Vielzahl an Maßen, Empfehlungen und insbesondere der Schwierigkeit, angemessene Effektstärken für die Korpuslinguistik zu definieren, wird in dieser Untersuchung der Empfehlung von Evert und Baroni gefolgt und für jeden Messwert ein 95%-Konfidenzintervall für die Differenz zwischen den jeweiligen Anteilen des untersuchten Merkmals in den miteinander verglichenen Datensätzen angegeben. Ein Konfidenzintervall beschreibt das Intervall, „[…] das den wahren aber unbekannten Parameter mit einer vorgegebenen Wahrscheinlichkeit […] überdeckt“ (Sachs/Hedderich 2009[13]: 291, sic!). Der Vorteil eines solchen Intervalls gegenüber konventionellen Maßzahlen für die Effektstärke besteht dabei in dessen leichter Interpretierbarkeit im Hinblick auf die numerische Größe des Effekts (des Unterschieds zwischen den Anteilswerten). Insbesondere kann eine untere Grenze für den Unterschied zwischen den Anteilen des interessierenden Merkmals an den jeweils untersuchten Datensätzen angegeben werden. Beispielsweise lässt sich aus dem von Evert und Baroni (2009: 793) ermittelten Konfidenzintervall KI=[0,03, 0,25] der Schluss ziehen, dass der Unterschied zwischen den Anteilen des interessierenden Merkmals an den jeweils untersuchten Datensätzen mindestens 3% betragen muss, wobei der Anteil aufgrund des positiven Vorzeichens der Werte im Intervall im ersten Datensatz größer ist als im zweiten. Zur Durchführung der Tests wird die Statistik-Software R[59] (R Development Core Team 2012) verwendet, die TTR und die Wortartenfrequenzen werden dagegen mittels einfacher Perl-Skripte erhoben.

Tabelle 8 gibt für jedes der „Babouk“-Korpora die TTR sowie die relative Frequenz von Adverbien und Nomina an. *n* bezeichnet dabei die Länge (in *tokens*) des jeweiligen Datensatzes nach Entfernung nicht-alphabetischer Zeichen. Für jedes der untersuchten Merkmale ist in den darunter liegenden Zeilen die Prüfgröße des Chi-Quadrat-Tests χ^2 sowie mit „KI“ das 95%-Konfidenzintervall[60] für die Größe des Unterschieds zwischen den beiden verglichenen Anteilen an den jeweiligen Korpora angegeben. Für die TTR sowie die Wortarten-Frequenzen sind die Vergleichswerte für das jeweilige Referenzkorpus hinter dem Gedankenstrich angegeben. Für jeden Test gilt, dass $p \leq 0{,}0001$.

[59] http://cran.r-project.org/doc/manuals/R-intro.pdf, Zugriff am 24.03.2014.

[60] Die Berechnung erfolgt mit Hilfe der R-Funktion „prop.test“.

	EKD (n=961.146)	**TKD** (n=5.523.745)	**EKR** (n=320.806)	**TKR** (n=1.043.627)
TTR	0,0736 / 0,0872	0,0418 / 0,0449	0,0602 / 0,0814	0,0558 / 0,0509
χ^2	1202,7590	672,0294	1093,1470	244,4759
KI	[-0,0144, -0,0128]	[-0,0034, -0,0029]	[-0,0224, -0,0199]	[0,0043, 0,0055]
ADV	0,0422 / 0,0688	0,0416 / 0,0686	0,0254 / 0,0470	0,0298 / 0,0439
χ^2	6505,5750	38.472,7700	2153,2570	2948,3350
KI	[-0,0273, -0,0260]	[-0,0272, -0,0267]	[-0,0226, -0,0207]	[-0,0147, -0,0137]
NN	0,3294 / 0,2398	0,3260 / 0,2400	0,4576 / 0,3158	0,4431 / 0,3521
χ^2	18.944,8100	100.715,8000	13.591,0900	18.049,0600
KI	[0,0883, 0,0909]	[0,0855, 0,0865]	[0,1394, 0,1441]	[0,0897, 0,0923]

Tabelle 8: Vergleichende statistische Charakterisierung der gesammelten Korpora

Hinsichtlich der TTR können die in Tabelle 8 zusammengefassten Daten die Annahme von Unterschieden zwischen den Referenzdaten und den „Babouk"-Korpora nicht bestätigen: Für das Deutsche fallen die Unterschiede nur schwach aus (entsprechend des Konfidenzintervalls ca. 1% für das EKD und für das TKD noch weniger) und dies obwohl hier randomisierte Sätze des NCL mit Lemmata aus zusammenhängenden Texten verglichen werden, d. h. bei der Verwendung zusammenhängender Texte als Referenzmaterial könnte sich der Vergleich sogar noch in die entgegengesetzte Richtung entwickeln. Für das Russische besteht zwar ein größerer Unterschied zwischen dem EKR und dem RIC, das TKR dagegen hat jedoch sogar eine höhere TTR als das Referenzkorpus. Eine Ursache hierfür könnten fehlerhafte Lemmatisierungen der zahlreichen russischen Wortformen sein, so dass Lemmata in unterschiedlichen Formen mehrfach gezählt werden, allerdings kann davon ausgegangen werden, dass derlei Fehler auch in den Referenzdaten auftreten (wobei freilich unklar ist, in welchem Maße). Zudem war bereits darauf hingewiesen worden, dass das RIC (im Unterschied zum NCL) einen mittleren Fachlichkeitsgrad aufweist. Die Ergebnisse für die TTR könnten ein Indiz dafür sein, dass sich die russischen „Babouk"-Korpora auf einem ähnlichen, nicht aber einem höheren Fachlichkeitsniveau bewegen. Eine andere Interpretation des Ergebnisses könnte so lauten, dass die TTR kein zuverlässiges quantitatives Maß für die Fachlichkeit eines Textes darstellt. Letztendlich kann diese Frage auf Grundlage der hier dargestellten Daten nicht beantwortet werden.

Relevante Unterschiede können dagegen hinsichtlich der relativen Frequenzen von Adverbien und Nomina ermittelt werden. Diese Unterschiede deuten im Sinne der Studie von Teich und Fankhauser (2010) auf einen für fachlich spezialisierte Texte typischen Nominalstil sowie einen tendenziellen Verzicht auf Modifikationen oder Relativierungen mittels Adverbien hin. Werden diese Merkmale im Sinne der zitierten Studie als Fachtextmerkmale interpretiert, kann von einem mittleren Fachlichkeitsniveau der „Babouk"-Korpora ausgegangen werden. Eingehendere Untersuchungen quantitativer Merkmale von Fachtexten sowie eine Abgrenzung der „Babouk"-Korpora nach oben, d. h. gegen stark fachliche Texte (wie z. B. die von Walter untersuchten Gerichtsurteile) könnten Gegenstand zukünftiger Studien sein.

Weitere interessante Maße für die Beschreibung der gesammelten Korpora sind die Anzahl der Sätze pro Korpus sowie die durchschnittliche Satzlänge. Tabelle 9 stellt diese Maße zusammenfassend dar.

Maß	EKD	TKD	EKR	TKR
Anzahl Sätze	64.438	538.465	35.507	111.214
Durchschnittliche Satzlänge	15,41	13,50	9,94	9,88

Tabelle 9: Wort- und Satzzahl in den Korpora sowie durchschnittliche Satzlänge

Die Angabe der Satzzahl pro Korpus resultiert aus der Ausgabe eines Satzgrenzenerkenners, der ein Korpus in einzelne Sätze zerlegt. Zur Satzgrenzenerkennung in den russischen und deutschen Korpora wird in vorliegender Arbeit das Perl-Modul Lingua::Sentence[61] verwandt, das in einem einfachen Perl-Skript implementiert wurde (s. Abschnitt 4.4). Zu betonen ist, dass ein solches Modul nicht fehlerfrei arbeitet, also irrtümliche Trennungen beispielsweise nach Abkürzungen vornimmt (falsche Positive) bzw. notwendige Trennungen nicht durchführt (falsche Negative). Die durchschnittliche Satzlänge für ein Korpus ergibt sich dann, wenn der Wortumfang eines Korpus' (hierfür werden die in Tabelle 6 dargestellten Werte *vor* der Satzgrenzenerkennung verwendet) zur Anzahl der Sätze im Korpus in Beziehung gesetzt wird. Angesichts der automatischen – und demnach unvollkommenen – Satzgrenzenerkennung sind die so erhaltenen Werte lediglich als Näherungswerte zu verstehen. Dennoch tragen die Ergebnisse der Satzgrenzenerkennung zu einer ersten Beschreibung der erstellten Korpora insofern bei, als sie Einblicke in Unterschiede in der Struktur der Korpora erlauben. So zeigt Tabelle 9 zunächst eine Differenz zwischen den Satzlängen der deutschen und russischen Korpora auf. Des Weiteren zeigt die Analyse einen recht deutlichen Unterschied zwischen dem EKD und dem TKD. Wird die von Baumann (1998: 732) für populärwissenschaftliche (deutsche) Texte genannte durchschnittliche Satzlänge von 20-22 Wörtern zugrunde gelegt, wird allerdings deutlich, dass die hier analysierten Texte – zumindest auf Grundlage der bisherigen Untersuchung – noch weit unter diesem Niveau liegen und möglicherweise ein noch geringeres Anspruchsniveau bedienen bzw. eine recht starke Durchmischung mit Texten anderer Textsorten abbilden.

4.1.4.2 Vergleich des Entwicklungskorpus' mit dem Testkorpus

Neben einem Vergleich der erstellten Korpora mit allgemeinsprachlichen Referenzkorpora, mit dessen Hilfe Indizien dafür gesammelt werden können, dass die Korpora der Anforderung einer fachlichen Spezialisierung genügen, ist auch der Vergleich der Entwicklungs- mit den Testkorpora von Interesse. Ziel eines solchen Vergleichs ist es, festzustellen, ob sich das für die jeweilige Sprache erstellte Korpuspaar thematisch voneinander so stark unterscheidet, dass das Testkorpus zur Überprüfung der Leistungsfähigkeit der zu entwickelnden Extraktionsmethode in einem anderen Fachgebiet verwendet werden kann.
Hierzu sollen lexikalische Kriterien herangezogen werden. Mit Hilfe des Werkzeugs AntConc[62] (Anthony 2005) können beispielsweise Schlüsselwörter eines Korpus' ermittelt werden. Dabei werden im Korpus vorkommende lexikalische Einheiten unter Zuhilfenahme statistischer Werkzeuge wie
log-likelihood ratio oder Chi-Quadrat mit den lexikalischen Einheiten in einem Referenzkorpus verglichen und diejenigen Einheiten herausgefiltert, die für das analysierte Korpus charakteristisch sind. Zunächst sollen die beiden russischen Korpora miteinander verglichen werden. Unter

[61] http://search.cpan.org/~achimru/Lingua-Sentence-1.05/lib/Lingua/Sentence.pm, Zugriff am 24.03.2014.
[62] http://www.antlab.sci.waseda.ac.jp/software.html, Zugriff am 24.03.2014.

Zuhilfenahme entsprechend großer Teile des RIC gibt AntConc für das EKR 1735 und für das TKR 5296 Schlüsselwörter aus. Tabelle 10 zeigt für beide Korpora eine um nicht-terminologische Elemente bereinigte Liste der jeweils zehn am höchsten gerankten einteiligen Schlüsselwörter mit ihren deutschen Übersetzungen.

EKR	**Übersetzung**	**TKR**	**Übersetzung**
двигатель	Motor	газ	Gas
автомобиль	Kraftfahrzeug	энергия	Energie
вал	Welle (Mechanik)	излучение	Strahlung
цилиндр	Zylinder	двигатель	Motor
тормоз	Bremse	топливо	Treibstoff
клапан	Ventil	метр	Meter
давление	Druck	взрыв	Explosion
колесо	Rad	давление	Druck
поршень	Kolben	нефть	Erdöl
миллиметр	Millimeter	мощность	Leistung

Tabelle 10: Top-10 terminologisch relevante Schlüsselwörter der russischen Korpora nach Analyse mit AntConc

Diese Liste der Schlüsselwörter scheint die in Tabelle 6 beschriebene inhaltliche Spezialisierung beider Korpora – für das EKR lautet das Fachgebiet KFZ, das TKR deckt dagegen mehrere Gebiete, insbesondere Energie und Kernenergie, KFZ und Physik ab – zu bestätigen. Wird nun nach terminologischen Überlappungen zwischen den Korpora gefragt, kann ein Vergleich der Schlüsselwortlisten aufschlussreich sein. Ein Vergleich der Schlüsselwortlisten für beide Korpora ergibt eine Überlappungsgröße von 1216 Einheiten. Wird diese Liste allerdings derart bereinigt, dass nur korrekt lemmatisierte, terminologisch relevante Nomina erfasst werden, bleiben lediglich 421 in beiden Korpora auftretende Schlüsselwörter übrig, von denen einige falsch lemmatisierten Formen derselben Grundform entsprechen. Dies entspricht der ebenfalls aus Tabelle 6 zu entnehmenden Tatsache, dass die beiden Korpora zum Teil dieselben sowie verwandte Themenbereiche abdecken. Um die Möglichkeit einer Überprüfung der Leistungsfähigkeit des zu erarbeitenden Extraktionsalgorithmus' auch in neuen Fachgebieten zu gewährleisten, wird bei der Erstellung des Goldstandards darauf zu achten sein, dass die Entwicklungs- und Testdaten möglichst keine Überschneidungen aufweisen.

Zur Analyse der beiden deutschen Korpora mit AntConc wurde wiederum ein Teil des NCL als Referenzkorpus verwendet. Die Analyse ergibt für das EKD 4929 und für das wesentlich größere TKD 15.129 Schlüsselwörter. Für die beiden deutschen Korpora gibt Tabelle 11 die um nicht terminologisch relevante Elemente bereinigten Listen der jeweils zehn am höchsten gerankten einteiligen, d. h. aus lediglich einem graphischen Wort bestehenden, Schlüsselwörter an.

EKD	**TKD**
Anlage	Energie
Windkraftanlage	Strom
Windenergie	Anlage
Energie	Photovoltaik
Wind	Kondensator
Offshore	Kraftwerk
Windpark	Wärme
Windenergieanlage	Spannung

Rotor	Biomasse
Rotorblatt	Akku

Tabelle 11: Top-10 terminologisch relevante Schlüsselwörter der deutschen Korpora nach Analyse mit AntConc

Das TKD scheint neben echten deutschen Termini auch eine große Anzahl englischer Ausdrücke aus dem Bereich IT – hier möglicherweise insbesondere aus zahlreichen Forumstexten – zu enthalten. Ein Vergleich der AntConc-Ergebnisse für beide Korpora erbringt eine Schlüsselwort-Übereinstimmung von 2565 lexikalischen Einheiten. Werden indes wieder lediglich die terminologisch relevanten Nomina berücksichtigt, reduziert sich die Anzahl der übereinstimmenden Schlüsselwörter auf 755 lexikalische Einheiten. Wie im Falle der russischen Korpora wird bei der Erstellung des Goldstandards darauf zu achten sein, dass Entwicklungs- und Testdaten keine Überschneidungen enthalten.

4.2 Annotationsstudie zur Erstellung der Datenbasis

In Abschnitt 3.4.4 waren unterschiedliche Aspekte der Evaluation von Extraktionsmethoden diskutiert worden. Das grundsätzliche Problem bei der Evaluierung von Methoden zur Extraktion wissenshaltiger Kontexte aus Textkorpora besteht dabei, wie bereits dargestellt, darin, dass – sollen die klassischen Evaluierungsmaße Präzision und Abdeckung angewendet werden – ein Maßstab für die optimale Leistung einer Extraktionsmethode – ein sogenannter Goldstandard – existieren muss. Aus den in Abschnitt 3.4.4 genannten Ansätzen zur Erstellung eines Goldstandards soll in vorliegender Arbeit der Weg einer Teilanalyse der in den vorstehenden Abschnitten beschriebenen Textkorpora begangen werden. Der Goldstandard wird dabei auf dem Wege der intersubjektiv überprüften Annotation valider KRCs für jedes Korpus gesondert erarbeitet. Ein Beispiel für ein solches Vorgehen ist etwa die Arbeit von Walter (2010). Anzumerken ist in diesem Zusammenhang, dass die Erstellung eines solchen Goldstandards mit Duplikatentfernung und mehrfacher Annotation eher als iterativer oder zyklischer Prozess zu charakterisieren ist, in dessen Verlauf Arbeits-, Überprüfungs- und Korrekturphasen einander abwechseln. Der leichteren Verständlichkeit wegen wird der Prozess im Folgenden jedoch sequentiell dargestellt.

4.2.1 Ausgangsdaten

Die Annotation der Daten wurde in zwei Schritten durchgeführt. Ein erster Satz von KRC-Kandidaten wurde von der Autorin vorliegender Arbeit (im Folgenden: Annotator A) als möglicherweise valide KRCs aus Unterabschnitten der vier vorliegenden Korpora ausgewählt. Dabei kamen als Verfeinerung der in Abschnitt 2.3.1 entwickelten Definition wissenshaltiger Kontexte sowie der in Abschnitt 2.3.3.2 erarbeiteten Liste akzeptabler semantischer Relationen die folgenden Validitätskriterien zum Einsatz:

- Der KRC-Kandidat kann einem Zielterminus zugeordnet werden.
- Die zeitliche Gültigkeit der durch den KRC-Kandidat vermittelten Information ist nachvollziehbar.
- Der KRC-Kandidat kann einem Fachgebiet zugeordnet werden.
- Die durch den KRC-Kandidaten vermittelte Information ist generalisierbar oder trägt einen relevanten Teilaspekt zum Verständnis des Begriffs bei.

- Der KRC-Kandidat enthält relevante Informationen zu Attributen oder semantischen Relationen.

Insbesondere aufgrund des Generalisierbarkeitskriteriums ergeben sich dabei bereits Folgen für die linguistische Gestalt wissenshaltiger Kontexte in den noch zu untersuchenden Daten, da beispielsweise die Verwendung von Eigennamen sowie bestimmte Formen der Pronominalisierung als Kriterien für einzelfallbezogene Informationen interpretiert wurden und Kandidaten mit diesen Merkmalen entsprechend nicht in die Annotationsdaten aufgenommen wurden. Die Auswahl von KRC-Kandidaten beschränkte sich zudem auf die Ebene des Satzes. Tabelle 12 gibt einen Überblick über die Anzahl der für alle vier Korpora von Annotator A ausgewählten KRC-Kandidaten für den ersten Annotationsschritt.

	EKR	**TKR**	**EKD**	**TKD**
Anzahl ausgewählter KRC-Kandidaten	299	229	372	299
KRC-Kandidaten pro Sprache		**528**		**671**

Tabelle 12: Überblick über die von Annotator A ausgewählten KRC-Kandidaten für den ersten Annotationsdurchgang

Damit standen für das Russische über 500 und für das Deutsche über 600 KRC-Kandidaten zur Analyse während des ersten Annotationsdurchgangs bereit. Die Datensätze enthielten dabei keine Repetitionen – mehrfach in einem Korpus auftretende KRC-Kandidaten wurden, soweit möglich, nur einmal erfasst bzw. Duplikate entfernt. Überschneidungen zwischen Test- und Entwicklungskorpora wurden weitestgehend vermieden. Angestrebt wurde bei der Auswahl der KRC-Kandidaten außerdem, dass die Testdatensätze die inhaltlich-thematische Vielfalt der Testkorpora zumindest in Ansätzen widerspiegeln, um auf diese Weise die Voraussetzung für eine Überprüfung der Leistungsfähigkeit der zu entwickelnden Extraktionsmethode in neuen Fachgebieten zu überprüfen.
Nach Auswertung des ersten Annotationsdurchgangs und erfolgter Auswahl von KRCs für den Goldstandard wurden Subkorpora von Texten erstellt, welche die Sätze des im ersten Annotationsschritt erstellten vorläufigen Goldstandards enthalten. Um eine möglichst eingehende Analyse der in diesen Subkorpora enthaltenen Daten zu gewährleisten, wurden diese Subkorpora erneut von Annotator A auf möglicherweise noch nicht erfasste KRC-Kandidaten überprüft. Diese Kandidaten wurden in einem zweiten Annotationsdurchgang bewertet. Tabelle 13 gibt einen Überblick über die Daten für den zweiten Annotationsdurchgang.

	Sub-EKR	**Sub-TKR**	**Sub-EKD**	**Sub-TKD**
Anzahl ausgewählter KRC-Kandidaten	73	105	80	55
KRC-Kandidaten pro Sprache		**178**		**135**

Tabelle 13: Überblick über die von Annotator A ausgewählten KRC-Kandidaten für den zweiten Annotationsdurchgang

Um die Konsistenz des Goldstandards zu gewährleisten, wurden schließlich in einem letzten Arbeitsschritt alle ausgewählten KRC-Kandidaten im Hinblick auf die oben genannten Kriterien von

Annotator A ein weiteres Mal auf ihre Übereinstimmung mit den entwickelten Validitätskriterien hin überprüft und zweifelhafte Daten aus dem vorläufigen Goldstandard entfernt.

4.2.2 Annotationsszenario und Annotation

Vier studentische Annotatoren – fortgeschrittene Studenten des Instituts für Angewandte Linguistik und Translatologie an der Universität Leipzig – wurden mit der Annotation der KRC-Kandidaten in ihrer jeweiligen Muttersprache (Deutsch bzw. Russisch) im ersten Annotationsdurchgang beauftragt. Die Annotationsaufgabe bestand darin, die in einer Tabelle – d. h. ohne zusätzliche Kontextinformationen – präsentierten Kandidaten im Fall von Validität mit 1 und im Fall von Invalidität mit 0 zu annotieren. Darüber hinaus waren die Annotatoren gebeten worden, für jeden als valide bewerteten Kandidaten die Konfidenz mit 2 für „sicher“ und 1 für „unsicher“ anzugeben. Zur Erleichterung der Aufgabe enthielt die zu annotierende Datei für jeden KRC-Kandidaten mindestens einen potentiellen Zielterminus, für den der Kandidat möglicherweise eine relevante Erklärung enthielt. Zieltermini waren vorab von Annotator A identifiziert worden, wobei auch in den KRC-Kandidaten auftretende Varianten berücksichtigt wurden. Im Falle des Auftretens mehrerer Definienda in einem KRC-Kandidaten wurde in der Regel nur ein Definiendum angegeben. Beispiele dafür, wie die KRC-Kandidaten den Annotatoren präsentiert wurden, finden sich in Tabelle 14.

Bewertung (1 = valide, 0 = nicht valide)	**Konfidenz (2 = sicher, 1 = unsicher; bei nicht validen Kandidaten kann die Angabe entfallen)**	**Zielterminus (Terminus/Termini, für die eine Erklärung geliefert wird)**	**Kontextkandidat**
		erneuerbare Energie	Als erneuerbare Energien werden Energiequellen bzw . Energieträger bezeichnet , die kurzfristig und nach menschlichen Maßstäben unerschöpflich zur Verfügung stehen .[63]
		Ammoniak	Ammoniak ist aus einzelnen Molekülen aufgebaut , die jeweils aus einem Stickstoff - und drei Wasserstoffatomen bestehen .[64]

Tabelle 14: Beispiel für die Präsentation der KRC-Kandidaten während der Annotation des Goldstandards

Um sicherzustellen, dass die Aufgabe von den Annotatoren richtig verstanden wurde, wurden vor Beginn der Erstannotation jedem der Annotatoren zunächst etwa 50 Sätze aus den

[63] sic!
[64] sic!

Entwicklungskorpora vorgelegt. Erst danach wurden – bei angemessener Annotation der Testsätze – die übrigen KRC-Kandidaten übergeben, andernfalls wurden weitere Annotationshinweise gegeben. Im Rahmen eines Einführungsseminars wurden die Annotatoren auf die Studie vorbereitet, indem ihnen die Validitätskriterien sowie der fachliche Hintergrund des Annotationsprojektes erläutert wurden[65], die in Kapitel 2 erarbeitete Definition wissenshaltiger Kontexte sowie die Validitätskriterien und eine Aufstellung der semantischen Zielrelationen lagen den Studenten schriftlich vor. Für den zweiten Annotationsdurchgang wurden vier erfahrene Übersetzer mit der Annotation beauftragt, die ebenfalls KRC-Kandidaten in ihrer Muttersprache annotierten. Auch ihnen lagen die Validitätskriterien und eine Aufstellung der semantischen Zielrelationen schriftlich vor. Auch im zweiten Annotationsschritt war zudem zu jedem KRC-Kandidat mindestens ein Zielterminus angegeben worden.

Anzumerken ist – Anhang 1 ist dies leicht zu entnehmen –, dass im Zusammenhang mit der Beschreibung der Annotationsaufgabe auf detaillierte Angaben zu linguistischen Eigenschaften wissenshaltiger Kontexte verzichtet wurde, da ja eine der Zielstellungen vorliegender Arbeit die umfangreiche linguistische Untersuchung wissenshaltiger Kontexte im Deutschen und Russischen ist. Einige der definitorischen Aussagen in der wissenschaftlichen Fachdiskussion häufig zugeschriebenen Eigenschaften sind bereits in Kapitel 3 kritisch diskutiert worden. Ebenfalls wurde bereits auf die Gefahr von Zirkelschlüssen hingewiesen. Nun ist freilich die Annotation von Elementen, deren Eigenschaften nicht konkret beschrieben werden können, alles andere als unproblematisch. Allerdings wurde bei der Konzeption der Annotationsschritte davon ausgegangen, dass die spezifische übersetzerische Kompetenz der Annotatoren bei der Auswahl von KRCs zum Tragen kommt. Entsprechend wurde die Annotationsaufgabe explizit in einem terminologischen Kontext präsentiert: Auszuwählen waren also solche Sätze, die bei der Bearbeitung eines terminologischen Problems von Nutzen sein könnten. Freilich wäre die Simulation eines solchen Problems anhand einer konkreten Übersetzungs- oder Rechercheaufgabe aus methodischer Sicht für das das hier verfolgte Forschungsinteresse noch vorteilhafter gewesen, die Konzeption und Auswertung einer solchen Studie war jedoch aufgrund beschränkter Ressourcen nicht zu bewältigen.

Aus dem übersetzungsbezogenen Annotationsszenario ergibt sich auch die Notwendigkeit einer Vorab-Annotation von Zieltermini/Definienda, da in einem solchen Szenario die zu bearbeitenden Termini in der Regel bekannt sind. Eine Verzerrung der Ergebnisse kann aufgrund dieser Vorgehensweise aber nicht angenommen werden, da diesem Arbeitsschritt eine auf Exhaustivität orientierte Auswahl von KRC-Kandidaten durch Annotator A vorausgegangen war und erst danach Zieltermini in den ausgewählten Kandidaten bestimmt wurden.

Mit der in Walter (2010: 72 ff.) beschriebenen Annotationsstudie – die umfangreichen Experimente Cramers (2011) können im Rahmen dieser Arbeit ohnehin nicht repliziert werden – ist die hier beschriebene Vorgehensweise mithin nicht direkt vergleichbar. Wurde in der Arbeit Walters ein insgesamt 7627 Sätze umfassendes Korpus deutscher Gerichtsentscheidungen von zwei Annotatoren – einem Linguisten und einem Juristen – jeweils vollständig durchgearbeitet, erschien ein solches Vorgehen für das hier verfolgte Forschungsvorhaben angesichts des Umfangs und der inhaltlichen Vielfalt der zur Verfügung stehenden Daten nicht umsetzbar.

[65] Die entsprechenden Power-Point-Folien sind in Anhang 1 zusammengefasst.

4.2.3 Ergebnisse und Bewertung

4.2.3.1 Deutsch

Die insgesamt 671 deutschen KRC-Kandidaten des ersten Annotationsdurchgangs wurden von den Annotatoren De1 und De2 – beide deutsche Muttersprachler – bearbeitet. Tabelle 15 zeigt die Ergebnisse der Annotation einschließlich der durchschnittlichen Konfidenz (über der Summe der von den Annotatoren als valide bewerteten Sätze) der Annotation im Überblick.

	De1		**De2**	
	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz
EKD	347 (93%)	1,66	341 (92%)	1,84
TKD	290 (97%)	1,70	263 (88%)	1,83

Tabelle 15: Ergebnisse der deutschen Erstannotation pro Annotator

Annotator De2 hat bei der Annotation der KRC-Kandidaten in Sub-EKD und Sub-TKD jeweils einen Satz übersehen – diese gingen als nicht valide Sätze in die Bewertung der Annotation ein. Auch bei der Bestimmung der Konfidenz der validen KRCs in dem Datensatz für die Testannotation traten bei beiden Annotatoren dahingehend Inkonsistenzen auf, als auch für invalide Sätze ein Konfidenzwert angegeben wurde. Nach der Besprechung der Testannotation trat dieser Fehler nicht mehr auf, die fälschlich angegebenen Werte wurden in 0 geändert. Der Vergleich der in Tabelle 15 angegebenen Daten für De1 und De2 zeigt auf den ersten Blick, dass De1 zwar insgesamt positiver, jedoch mit deutlich geringerer Konfidenz annotiert hat. In diesem Zusammenhang ist jedoch festzuhalten, dass die Konfidenz auch für De1 sowohl für das Sub-EKD als auch das Sub-TKD deutlich über dem Mittelwert von 1,5 (bei einer binären Entscheidung mit 1 = unsicher und 2 = sicher) liegt und damit als akzeptabel angesehen werden kann. Tabelle 16 gibt einen genaueren Überblick über die Annotation von De1 und De2 in Form einer Vierfeldertafel:

EKD					**TKD**				
	De1					De1			
De2		1	0	Gesamt	De2		1	0	Gesamt
	1	321	20	341		1	258	5	263
	0	26	5	31		0	32	4	36
	Gesamt	347	25	**372**		Gesamt	290	9	**299**

Tabelle 16: Vierfeldertafel für die Erstannotation der deutschen Korpora

Die für den zweiten Annotationsschritt ausgewählten 135 KRC-Kandidaten wurden von den Annotatoren De3 und De4 – erfahrenen Übersetzern und wiederum deutschen Muttersprachlern – bearbeitet. Auch hier traten in der Annotation dahingehend Inkonsistenzen auf, dass Annotator De3 auch für negativ bewertete Kandidaten einen Konfidenzwert angab. Um die Vergleichbarkeit der Ergebnisse zu gewährleisten, wurden diese Werte in 0 geändert. Sowohl Annotator De3 als auch Annotator De4 vergaben für jeweils einen positiv bewerteten Satz aus dem Sub-EKD einen Konfidenz-Wert von 0, der in der Auswertung beibehalten wurde. Tabelle 17 gibt einen Überblick über die Annotationsergebnisse für die Annotatoren De3 und De4.

	De3		**De4**	
	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz
Sub-EKD	63 (79%)	1,71	66 (83%)	1,50
Sub-TKD	45 (82%)	1,53	45 (82%)	1,51

Tabelle 17: Ergebnisse der deutschen Zweitannotation pro Annotator

Für diesen Annotationsdurchgang auffällig ist der beträchtliche Unterschied zwischen den Konfidenzwerten von Annotator De3 bei ähnlichem Anteil positiver Annotationen, der möglicherweise die aktive Reflektion über die gestellte Annotationsaufgabe widerspiegelt. Annotator De4 dagegen annotierte durchgängig mit Konfidenzwerten, die dicht an dem hier als Mindestwert angesehenen Mittelwert von 1,5 liegen. Tabelle 18 stellt die Ergebnisse für diesen Annotationsschritt in Form einer Vierdfeldertafel dar.

Sub-EKD					**Sub-TKD**				
	De3					De3			
De4		1	0	Gesamt	De4		1	0	Gesamt
	1	50	16	66		1	38	7	45
	0	13	1	14		0	7	3	10
	Gesamt	63	17	**80**		Gesamt	45	10	**55**

Tabelle 18: Vierfeldertafel für die Zweitannotation der deutschen Korpora

Die Tabellen zeigen wie schon für den ersten Annotationsdurchgang relativ große Diskrepanzen zwischen den Bewertungen der beiden Annotatoren auf. Zwar treten eindeutig negative Bewertungen selten auf, jedoch können die häufigen voneinander abweichenden Bewertungen als Hinweis auf einen Grenzbereich der Kategorie des wissenshaltigen Kontexts interpretiert werden, für den unterschiedliche Nutzer zu individuellen Bewertungen gelangen und der sich methodisch nur schwer erfassen lässt.

4.2.3.2 Russisch

Die insgesamt 527 russischen KRC-Kandidaten wurden im ersten Annotationsdurchgang von den Annotatoren Ru1 und Ru2 – beide russische Muttersprachler – bearbeitet. Tabelle 19 zeigt die Ergebnisse der Annotation einschließlich der durchschnittlichen Konfidenz (über der Summe der von den Annotatoren als valide bewerteten Sätze) der Annotation im Überblick.

	Ru1		**Ru2**	
	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz
Sub-EKR	289 (97%)	1,98	294 (98%)	1,83
Sub-TKR	229 (100%)	1,99	225 (98%)	1,85

Tabelle 19: Ergebnisse der russischen Erstannotation pro Annotator

Tabelle 20 gibt einen detaillierteren Überblick über die Annotation der russischen Korpora durch Ru1 und Ru2.

Sub-EKR					**Sub-TKR**				
	Ru1					Ru1			
Ru2		1	0	Gesamt	Ru2		1	0	Gesamt
	1	287	7	294		1	225	0	225
	0	2	3	5		0	4	0	4
	Gesamt	289	10	**299**		Gesamt	229	0	**229**

Tabelle 20: Vierfeldertafel für die Erstannotation der russischen Korpora

Die Zweitannotation der russischen Korpora wurde durch erfahrene Übersetzer – wiederum russische Muttersprachler – durchgeführt und ergab die in Tabelle 21 dargestellten Ergebnisse.

	Ru3		**Ru4**	
	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz	Anteil der als korrekt annotierten Sätze	Durchschnittliche Konfidenz
Sub-EKR	64 (88%)	1,80	64 (88%)	1,59
Sub-TKR	99 (94%)	1,86	102 (97%)	1,75

Tabelle 21: Ergebnisse der russischen Zweitannotation pro Annotator

Auch hier werden Unterschiede in der Annotationsstrategie der beiden Annotatoren deutlich. Die gestiegene Konfidenz von Annotator Ru4 bei der Annotation des russischen Testkorpus‘ im Vergleich zum russischen Entwicklungskorpus kann möglicherweise ebenfalls als Resultat eines aktiven Reflexionsprozesses interpretiert werden, in dessen Ergebnis – anders als bei Annotator De4 – eine größere Selbstsicherheit bei der Annotation entwickelt wurde. Ein anderer Erklärungsansatz wäre derjenige, dass die Qualität der KRC-Kandidaten für das Sub-TKR höher war und entsprechend in einer höheren Konfidenz und einem größeren Anteil positiver Annotation resultierte. Letztlich können diese Fragen aufgrund der erhobenen Daten aber nicht entschieden werden, sondern müssen an dieser Stelle in den Gegenstandsbereich möglicher zukünftiger empirischer Untersuchungen verwiesen werden. Auch ein eindeutiger Zusammenhang zwischen Konfidenz und KRC-Qualität scheint aus den gewonnenen Daten angesichts der unterschiedlichen Annotationsstrategien der Annotatoren nicht ableitbar. Tabelle 22 fasst die Ergebnisse der russischen Zweitannotation wiederum in Form von Vierfeldertafeln zusammen.

Sub-EKR					**Sub-TKR**				
	Ru3					Ru3			
Ru4		1	0	Gesamt	Ru4		1	0	Gesamt
	1	60	4	64		1	97	5	102
	0	4	5	9		0	2	1	3
	Gesamt	64	9	**73**		Gesamt	99	6	**105**

Tabelle 22: Vierfeldertafel für die Zweitannotation der russischen Korpora

4.2.3.3 Quantitative Bewertung der Annotation

In Abschnitt 3.3.2 waren Annotationsmethodologien ebenso wie Ansätze zur Bewertung von Annotationsexperimenten diskutiert worden. Aus den zur Verfügung stehenden Möglichkeiten für letzteren Aspekt soll hier das Verfahren der Bewertung mit Präzision, Abdeckung und f-Maß gewählt werden. Hierfür sprechen mehrere Gründe:

- Aufgrund der unterschiedlichen Annotationsszenarien sowie wegen der bereits diskutierten Abhängigkeit von κ von den Randsummen der Vierfeldertafeln ist die hier durchgeführte Annotationsstudie kaum mit den Ergebnissen anderer Studien vergleichbar. Ein weiteres wesentliches Argument ist der Umstand, dass die Annotatoren aufgrund von Ressourcenbeschränkungen nicht ganze Korpusabschnitte durchgehen konnten, sondern lediglich vorab ausgewählte KRC-Kandidaten bewerteten. Insbesondere wurden ihnen also keine eindeutig negativen Kandidaten vorgelegt. Das Fehlen einer breit vertretenen negativen Kategorie kann jedoch den κ-Wert verzerren.
- In diesem Zusammenhang ist auch zu beachten, dass die hier beschriebene Annotationsaufgabe auf Satzebene in Bezug auf die Annotationskomplexität nicht direkt mit weniger komplexen Annotationsaufgaben auf Wortebene – z. B. der Zuordnung von Wörtern zu Wortarten – vergleichbar ist und auch die Hypothese einer niedrigen Übereinstimmung zwischen den Annotatoren rechtfertigt. Zwar scheint die Annahme einer völlig ungerichteten Annotation angesichts der zielgerichteten Auswahl und Vorbereitung der Annotatoren nicht gerechtfertigt und widerspricht auch den Forschungsergebnissen etwa Cramers (2011), bei der insgesamt geringen Anzahl an Annotatoren können Fehlannotationen jedoch gravierende Auswirkungen haben. Das Annotationsergebnis erscheint mithin als nicht statistisch verallgemeinerbar, so dass angesichts einer derart komplexen Kategorie wie derjenigen des wissenshaltigen Kontexts aus individuellen Abweichungen der Annotatoren keine statistisch validen Rückschlüsse auf die Gültigkeit oder Ungültigkeit eines Kontextkandidaten gezogen werden, die individuellen Bewertungen vielmehr lediglich als Indizien für die Gültigkeit bzw. Ungültigkeit von Kontextkandidaten angesehen werden können.
- Die Bewertung der Ergebnisse durch die Größen Präzision, Abdeckung und f-Maß bietet vor diesem Hintergrund immerhin die Möglichkeit zu überprüfen, wie weitgehend die Annotatoren in den unterschiedlichen Annotationsstufen miteinander übereinstimmen, ohne dabei allerdings – im Unterschied zu κ (vgl. Abschnitt 3.3.2.2) – von einer zufällig erreichbaren Minimalübereinstimmung auszugehen, die aus den genannten Gründen hier nicht berechenbar scheint.

Die Berechnung der Evaluationsmaße soll am Beispiel der Vierfeldertafel für die Erstannotation der deutschen Korpora (Tabelle 16) kurz erläutert werden:

- Die Annotation von Annotator De1 wird als vorläufiger Goldstandard angenommen.
- Die Tabelle zeigt, dass De2 für das EKD insgesamt 341 KRC-Kandidaten als valide annotiert hat, wobei mit Annotator De1 eine Übereinstimmung in 321 Fällen besteht (nach der Darstellung in 12) Feld O_{11}). 31 KRC-Kandidaten wurden von De2 als negativ bewertet, wobei hinsichtlich der Negativ-Kategorie lediglich in fünf Fällen eine Übereinstimmung mit De1 festgestellt werden kann (nach der Darstellung in 12) Feld O_{22}).
- Die Präzision p_{De1} für Annotator De2 ergibt sich dann wie folgt:

13) $$p_{De1} = \frac{tp}{tp+fp} = \frac{321}{321+20} \approx 0{,}94$$

- Die Abdeckung r_{De1} wird dann wie folgt berechnet:

14) $$r_{De1} = \frac{tp}{tp+fn} = \frac{321}{321+26} \approx 0{,}93$$

Hiermit ergibt sich für diesen Teil der Gesamtannotation (bei einem Gewichtungsfaktor β = 1) ein f-Maß von 0,93. Tabelle 23 fasst die Evaluationsmaße für alle weiteren Teile der Gesamtannotation zusammen.

	EKD	**TKD**	**EKR**	**TKR**
Erstannotation				
Präzision	0,94	0,98	0,98	1
Abdeckung	0,93	0,89	0,99	0,98
f-Maß	*0,93*	*0,93*	*0,98*	*0,99*
Zweitannotation				
Präzision	0,76	0,84	0,94	0,95
Abdeckung	0,79	0,84	0,94	0,98
f-Maß	*0,77*	*0,84*	*0,94*	*0,96*

Tabelle 23: Evaluation der Annotation

Wie können diese Werte nun interpretiert werden? Die in der Tabelle dargestellten Werte zeigen Übereinstimmungen auf, die zumindest über ein rein zufälliges Übereinstimmungsniveau hinauszugehen scheinen. Inwieweit hier normative Effekte aufgrund der Tatsache, dass KRC-Kandidaten von der Versuchsleiterin ausgewählt wurden, eine Rolle spielen, kann freilich nicht abgeschätzt werden. Festzustellen ist zudem, dass, obwohl alle Annotatoren das Gros der ihnen präsentierten KRC-Kandidaten als „valide“ markiert hat, die stark unterschiedlichen Konfidenzwerte sowie die aus Tabelle 23 ersichtlichen, zumindest für die deutsche Zweitannotation recht starken Abweichungen für die Ausbildung unterschiedlicher Annotationsstrategien der einzelnen Annotatoren sprechen, wie sie auch von Cramer (2011) postuliert wurden. Gemeint ist damit, dass angesichts von Grenzfällen eher positiv oder eher negativ, dafür aber möglicherweise mit einer höheren bzw. niedrigeren Konfidenz annotiert wird. Die durchgängig höheren Werte für den ersten Annotationszyklus könnten zudem darauf zurückzuführen sein, dass in Vorbereitung der Zweitannotation ein besonderer Fokus auf Exhaustivität gelegt wurde, so dass möglicherweise mehr zweifelhafte KRC-Kandidaten in die Annotationsdaten aufgenommen wurden. Andere Unterschiede zwischen den einzelnen Datensätzen könnten auf Unterschiede zwischen den Annotatoren (Studenten vs. erfahrene Übersetzer) oder den annotierten KRC-Kandidaten (z. B. stark technisches vs. weniger technisches Fachgebiet) zurückzuführen sein, ohne eingehendere Untersuchungen können jedoch diesbezüglich keine sicheren Aussagen getroffen werden. Angesichts der geringen Zahl von Annotatoren können die hier ermittelten Werte, wie bereits diskutiert, nicht verallgemeinert werden. Eingehendere Studien mit konkreteren Annotationsvorgaben könnten die Aufgabe verfolgen, ein angemesseneres Bewertungsinstrumentarium für wissenshaltige Kontexte zu entwickeln.

4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards

Aufgrund der im vorigen Abschnitt dargestellten Schwierigkeiten bei der quantitativen Auswertung der Annotation verfolgt vorliegende Studie eine qualitative Strategie zur Konsolidierung des Goldstandards. Dabei werden zunächst alle von allen drei Annotatoren im ersten Annotationsdurchgang – Annotator A sowie De1 und De2 bzw. Ru1 und Ru2 – übereinstimmend als valide bewerteten Sätze als zum Kernbereich der Kategorie des wissenshaltigen Kontexts zugehörig betrachtet (Erstellung der vorläufigen Grundgesamtheit). Eine qualitative Analyse der von einem

oder mehreren Annotatoren als negativ bewerteten KRC-Kandidaten soll alsdann Entscheidungen über möglicherweise fälschlich als negativ annotierte und mithin dem Goldstandard ebenfalls hinzuzufügende Sätze ermöglichen (Identifizierung der falschen Negative). Dieses Verfahren wird für die Zweitannotation wiederholt. Um die Konsistenz des zu erstellenden Goldstandards zu gewährleisten, erscheint indes auch eine qualitative Analyse der von allen Annotatoren als positiv bewerteten KRC-Kandidaten angebracht. Hierbei werden die in die vorläufige Grundgesamtheit übernommenen KRC-Kandidaten ein weiteres Mal zu den entwickelten Validitätskriterien in Beziehung gesetzt und mögliche falsche Positive, d.h. fälschlich als valide annotierte Sätze, aus dem Goldstandard entfernt. Die Erstellung des Goldstandards verläuft also in drei Stufen, die in Abbildung 8 dargestellt werden.

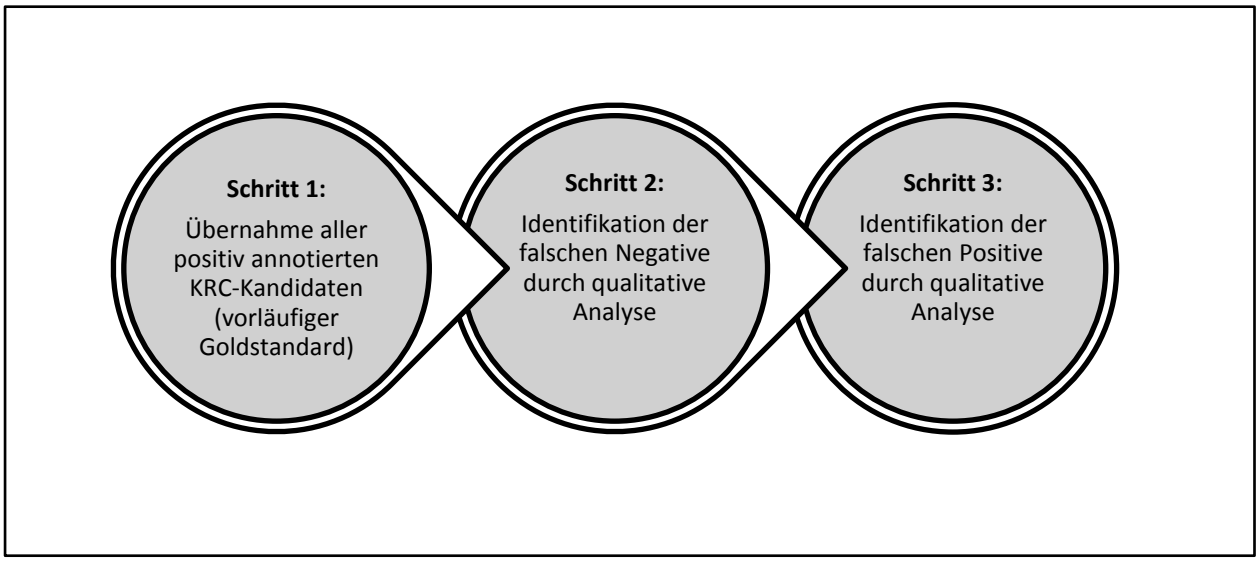


Abbildung 8: Arbeitsschritte zur Erstellung des Goldstandards (qualitative Bewertung)

4.2.3.4.1 Identifikation der falschen Negative

Tabelle 24 gibt einen Überblick über die Ergebnisse der Analyse negativ bewerteter deutscher Sätze für den ersten Annotationsdurchgang, indem sie negative Bewertungen, die entweder von beiden oder aber von einem der Annotatoren vorgenommen wurden, bestimmten Problemtypen zuordnet.

Problemtyp	**Sub-EKD**				**Sub-TKD**			
	beide	De1	De2	**Gesamt**	beide	De1	De2	**Gesamt**
Fachlicher Bezug unklar	1	1	1	3	0	0	3	3
Fehlannotation	*0*	*4*	*9*	*13*	*2*	*3*	*17*	*22*
Fehlerhafte Information	0	0	1	1	0	0	0	0
Information zu allgemein	1	0	4	5	0	0	2	2
Mehrsatz-Kontext	0	0	0	0	0	0	0	0
Prozess unklar	0	1	0	1	1	0	2	3
Semantische Relation	0	9	2	11	0	2	5	7

unklar								
Sprachliche Störung	0	0	0	0	0	0	0	0
Terminus unklar	1	2	7	10	0	0	2	2
Untergeordnete Bedeutung der terminologischen Information	2	2	0	4	1	0	0	1
Verallgemeinerbarkeit fraglich	0	0	2	2	0	0	0	0
Wörterbuchform	0	1	0	1	0	0	1	1
Gesamt	5	20	26	**51**	4	5	32	**41**

Tabelle 24: Ergebnis der qualitativen Analyse negativer Bewertungen für deutsche KRC-Kandidaten der Erstannotation

Wie Tabelle 24 zu entnehmen ist, wurden die negativen Bewertungen der Annotatoren De1 und De2 zu möglichen Ursachen für die negative Bewertung in Beziehung gesetzt. Die Kategorie „Fachlicher Bezug unklar“ bezeichnet dabei KRC-Kandidaten, die offenbar ohne zusätzliche Informationen keinem Wissensgebiet oder fachlich spezialisierten Zusammenhang eindeutig zuordenbar sind, so dass ein wesentliches Validitätskriterium verletzt wird. Ein Beispiel hierfür ist der Satz: „Auf den Etagenböden 7 sind die Flügel 1, 2 und 3 je Etage angeordnet.“ Die Kategorie „fehlerhafte Information“ bezeichnet KRC-Kandidaten, die inhaltliche Unzulänglichkeiten aufweisen, z. B.: „Darrieus-Blätter funktionieren gemäß dem Prinzip der Antriebskraft.“ Die Kategorie „Information zu allgemein“ bezeichnet Sätze mit relativ niedrigem Informationsgehalt, welche das Relevanzkriterium verletzen, etwa: „Auch die Nutzung der Windenergie gehört zu den schon lange bekannten Formen der Energieumwandlung.“ Mehrsatz-Kontexte sind Begriffserklärungen, die sich über mehrere Sätze erstrecken und aus pragmatischen Gründen aus der Untersuchung ausgeschlossen wurden. „Prozess unklar“ umfasst KRC-Kandidaten, die offenbar keinem terminologisch relevanten fachlichen Zusammenhang zugeordnet werden können, beispielsweise den Satz: „Der Läufer erzeugt also ein magnetisches Gleichfeld.“ Mitunter scheinen auch die dargestellten semantischen Relationen unklar zu sein, so dass nicht einzuschätzen ist, in welchem terminologisch relevanten Zusammenhang die durch den Kontext vermittelten Informationen zum Zielterminus stehen, etwa im Fall des Satzes: „Windkraftanlagen mit aktiver Stallregelung sind der Versuch, das Konzept der Stallregelung und des netzsynchronen Betriebs ohne teureren Gleich- und Wechselrichter auch auf größere Anlagen bis in den Megawattbereich zu übertragen.“ Sprachliche Störungen in der Satzoberfläche eines Kontext-Kandidaten werden in denjenigen Fällen als Argument für die Invalidität eines Kandidaten akzeptiert, wenn die Störung das Verständnis des terminologisch relevanten Zusammenhangs stark erschwert oder aber unmöglich macht, z. B. In diesem Satz: „Nun besteht aber eine Windkraftanlage, die zur Stromerzeugung dient aus drei wesentlichen Bausteinen, dem Windrotor, dem Getriebe und dem Generator, doch auch eine Turmkonstruktion und eine Steuerung der Anlage die aber auch zur Aufnahme und Speicherung der erzeugten elektrischen Energie dient.” Sätze der Kategorie „Terminus unklar“ scheinen keine eindeutige Identifikation eines Zielterminus’ zu erlauben oder enthalten einen fehlerhaften – z. B. falsch geschriebenen – Zielterminus, so dass ein weiteres Validitätskriterium verletzt wird, etwa im Fall des Satzes: „Auch die Entwicklung des finnischen Kapitänleutnants Sigurd Savonius gehört zu den Vertikalachskonvertern.“ Von De1 und De2 weitgehend übereinstimmend bewertet wurden KRC-Kandidaten, die offenbar keinem wissensvermittelnden Kontext entstammen und terminologisch relevante Information nur am Rande

vermitteln, beispielsweise der Satz: „Nimmt man die geplanten Windenergieprojekte auf See hinzu (Offshore-Windparks), so wird diese Betrachtung noch günstiger ausfallen.“ Wiederum für andere Sätze schien das Verallgemeinerbarkeitskriterium nicht erfüllt, so beispielsweise für: „Der Kern dieser Rotorblätter besteht aus einem ovalen Holm.“ Nicht zuletzt negativ bewertet wurden auch – etwas vereinfachend und von der tatsächlichen Form von Wörterbuchartikeln abstrahierend – solche Kandidaten, die eine an Wörterbuchartikel erinnernde Form – also ein vorangestelltes Lemma mit nachfolgender Erklärung (vgl. Pearson 1998: 162 ff.) – aufweisen, so beispielsweise der Satz: „Monopile (Offshore, pile: englisch für Pfahl, Pfosten) Dabei wird ein einzelner Mast im Erd- bzw. Seeboden versenkt.“ Oft handelt es sich hierbei um Strukturen, die außerhalb eines Wörterbuches oder Lexikons als ungrammatisch angesehen würden (ebd. 165) bzw. die nur sehr einfache Erklärungen für das vorangestellte Bezugswort liefern. Als Fehlannotationen werden dagegen solche Sätze angesehen, die fälschlich mit einer negativen Bewertung versehen wurden, d. h. solche Sätze, welche die in Abschnitt 4.2.1 genannten Validitätskriterien erfüllen und daher keinem der in Tabelle 24 genannten Negativitätskriterien zugeordnet werden können. Einen Überblick über Fehlannotationen in den deutschen Korpora vermittelt Anhang 2.

Tabelle 24 zeigt, dass Annotator De2, der anteilig mehr negative Bewertungen abgegeben hat, relativ viele invalide bzw. zum Grenzbereich der Kategorie des wissenshaltigen Kontexts zugehörige Sätze identifiziert, allerdings auch eine relativ hohe Zahl an Fehlannotationen vorgenommen hat. Deren Anteil bewegt sich indes im Verhältnis zu den insgesamt zu annotierenden KRC-Kandidaten noch auf einem akzeptablen Niveau, so dass die Bewertungen von De2 insgesamt als gültig angesehen werden können. Tabelle 25 fasst die Ergebnisse der qualitativen Analyse von Negativbewertungen für die Zweitannotation zusammen, wobei im Unterschied zu Tabelle 24 nur diejenigen Fehlerkategorien dargestellt werden, für die in diesem Annotationsdurchgang auch Beispiele gefunden wurden.

Problemtyp	**Sub-EKD**				**Sub-TKD**			
	beide	De3	De4	**Gesamt**	beide	De3	De4	**Gesamt**
Fachlicher Bezug unklar	1	7	0	8	0	0	1	1
Fehlannotation	*0*	*1*	*4*	*5*	*0*	*2*	*4*	*6*
Information zu allgemein	0	3	4	7	3	5	1	9
Mehrsatz-Kontext	0	0	1	1	0	0	0	0
Semantische Relation unklar	0	5	3	8	0	0	1	1
Sprachliche Störung	0	0	1	1	0	0	0	0
Gesamt	1	16	13	**30**	3	7	7	**17**

Tabelle 25: Ergebnis der qualitativen Analyse negativer Bewertungen für deutsche KRC-Kandidaten der Zweitannotation

Auch hier fallen Unterschiede in der Bewältigung der Annotationsaufgabe durch die Annotatoren ins Auge. So hat insbesondere Annotator De3 eine große Anzahl invalider KRC-Kandidaten identifiziert – ein Reflektionsprozess, der sich, wie ausgeführt, möglicherweise auch in der Entwicklung der Anotationskonfidenz dieses Annotators sowie in seinen – allerdings in vorliegender Arbeit nicht präsentierten – zahlreichen Anmerkungen und Kommentaren widerspiegelt. Für Annotator De4 fällt wiederum ein recht hoher Anteil an Fehlannotationen auf, der sich allerdings wie im Fall der

Erstannotation im Verhältnis zum Gesamtvolumen der Studie noch in einem akzeptablen Rahmen bewegt.
Tabelle 26 präsentiert die Ergebnisse der qualitativen Bewertung russischer negativ bewerteter Sätze aus der Erstannotation wiederum nur für diejenigen Kategorien, für die in dem untersuchten Datensatz auch tatsächlich Beispiele beobachtet wurden. Anhang 3 gibt darüber hinaus einen Überblick über negative Fehlannotationen der russischen KRC-Kandidaten.

Ursache	**Sub-EKR**				**Sub-TKR**			
	beide	Ru1	Ru2	**Gesamt**	beide	Ru1	Ru2	**Gesamt**
Fehlannotation	*2*	*7*	*0*	*9*	*0*	*0*	*0*	*0*
Information zu allgemein	0	0	0	0	0	0	3	3
Semantische Relation unklar	1	0	0	1	0	0	0	0
Sprachliche Störung	0	0	1	1	0	0	0	0
Verallgemeinerbarkeit fraglich	0	0	1	1	0	0	0	0
Gesamt	3	7	2	**12**	0	0	4	**3**

Tabelle 26: Ergebnis der qualitativen Analyse negativer Bewertungen für russische KRC-Kandidaten der Erstannotation

Die für die russischen KRC-Kandidaten abgegebenen negativen Bewertungen sind zahlenmäßig deutlich geringer und weisen auch eine andere Struktur auf. Es fällt auf, dass insbesondere Annotator Ru1 geringfügige sprachliche Störungen negativ bewertet hat. Allerdings sind linguistische Unregelmäßigkeiten nicht immer ein Indiz für eine schlechte Textqualität, sondern können beispielsweise auch auf Vorverarbeitungsfehler – etwa beim Umwandeln von PDF-Dokumenten in das txt-Format durch „Babouk“ – zurückzuführen sein. Darüber hinaus weist auch die Annotation von Annotator Ru1 keine Konsistenz in der Bewertung sprachlicher Störungen auf, eine generelle Ablehnung von KRC-Kandidaten mit oberflächlichen sprachlichen Unregelmäßigkeiten kann also nicht konstatiert werden. Obwohl sprachliche Korrektheit gerade in der Terminologie von eminenter Wichtigkeit ist, enthalten sprachlich leicht defekte Sätze unter Umständen terminologisch relevantes Wissen und sollten nicht ungeprüft entfernt werden. Die von Annotator Ru1 vorgenommene negative Bewertung von Kontext-Kandidaten mit geringfügigen Oberflächenstörungen wurde daher als Fehlannotation gewertet. Tabelle 27 schließlich präsentiert die qualitative Auswertung negativer Annotation für den zweiten russischen Annotationsdurchgang.

Ursache	**Sub-EKR**				**Sub-TKR**			
	beide	Ru3	Ru4	**Gesamt**	beide	Ru3	Ru4	**Gesamt**
Fehlannotation	*0*	*0*	*1*	*1*	*0*	*3*	*2*	*5*
Information zu allgemein	4	3	0	7	1	1	0	2
Semantische Relation unklar	1	0	2	3	0	0	0	0
Terminus unklar	0	0	1	1	0	0	0	0

Untergeordnete Bedeutung der terminologischen Information	0	1	0	1	0	0	0	0
Wörterbuchform	0	0	0	0	0	1	0	1
Gesamt	5	4	4	**13**	**1**	**5**	**2**	**8**

Tabelle 27: Ergebnis der qualitativen Analyse negativer Bewertungen für russische KRC-Kandidaten der Zweitannotation

4.2.3.4.2 Identifikation der falschen Positive

Das Vorliegen eines falschen Positivs kann dann konstatiert werden, wenn der betreffende KRC-Kandidat trotz übereinstimmend positiver Annotation einem der im vorigen Abschnitt entwickelten Negativitätskriterien zugeordnet werden kann. Zusätzlich zu den bereits genannten Kriterien kommen hier noch die Kriterien „Unvollständigkeit der Information“ und „Zeitliche Gültigkeit unklar“ ins Spiel. Unvollständige Informationen liegen dann vor, wenn offensichtlich relevante Informationsbestandteile im Satz nicht enthalten sind. Als unklar werden zeitbezogene Gültigkeitsangaben bewertet, wenn der Zeitraum einer eingeschränkten Gültigkeit der durch den Kontext-Kandidaten vermittelten Information nicht explizit angegeben ist. Innerhalb eines oder mehrerer Datensätze doppelt annotierte Sätze werden ebenfalls entfernt. Tabelle 28 gibt einen Überblick über das Ergebnis dieses Analyseschritts für die deutschen und russischen Korpora. Anhang 4 stellt die so entfernten deutschen Sätze im Überblick dar, Anhang 5 enthält die entsprechenden russischen Daten.

Ursache	**Sub-EKR**	**Sub-TKR**	**Sub-EKD**	**Sub-TKD**
Dopplung	9	0	1	0
Fachlicher Bezug unklar	14	12	7	10
Fehlerhafte Information	1	0	1	0
Information unvollständig	0	0	1	0
Information zu allgemein	20	12	20	8
Mehrsatz-Kontext	0	0	1	2
Prozess unklar	1	7	0	0
Sprachliche Störung	1	0	0	0
Semantische Relation unklar	2	3	7	2
Terminus unklar	8	4	2	4
Untergeordnete Bedeutung der terminologischen Information	0	0	0	0
Verallgemeinerbarkeit fraglich	8	12	10	1
Wörterbuchform	1	9	1	2
Zeitliche Gültigkeit unklar	0	0	1	0
Gesamt	**65**	**59**	**52**	**29**

Tabelle 28: Ergebnis der qualitativen Analyse positiver Bewertungen für die untersuchten Korpora

4.2.3.4.3 Goldstandard

Für die Erstellung des Goldstandards wird nun die Strategie gewählt, dass zunächst diejenigen Sätze übernommen werden, die von allen Annotatoren als gültig bewertet wurden. Alsdann werden die irrtümlich von den Annotatoren als falsch bewerteten Sätze (negative Fehlannotationen) ebenfalls übernommen. Positive Fehlannotationen werden schließlich aus dem Goldstandard entfernt. Tabelle 29 stellt die sich ergebenden Größen des Goldstandards im Überblick dar.

	Sub-EKR	**Sub-TKR**	**Sub-EKD**	**Sub-TKD**
Erstannotation				
Von allen Annotatoren positiv bewertete Sätze	287	225	321	258
Negative Fehlannotationen	9	0	13	22
Zweitannotation				
Von allen Annotatoren positiv bewertete Sätze	60	97	50	38
Negative Fehlannotationen	1	5	5	6
Positive Fehlannotationen	-65	-59	-52	-29
Gesamt	**292**	**268**	**337**	**295**
KRCs pro Sprache		**560**		**632**

Tabelle 29: Größe des Goldstandards nach Erst- und Zweitannotation sowie der qualitativen Konsolidierung

4.2.4 Abdeckung der Annotation

Nach dem Abschluss der Annotation soll als qualitätssichernde Maßnahme eine Abschätzung der Abdeckung der Annotation durchgeführt werden, d. h. anhand einer einfachen Zufallsstichprobe (vgl. Mosler/Schmid 2006[2]: 175, 182) aus jedem der erstellten Subkorpora soll abgeschätzt werden, wie viele relevante KRCs bei der Annotation übersehen wurden. Dies scheint erforderlich, da auch das zweimalige Durchgehen von Annotator A durch die Texte der Subkorpora zur Auswahl von KRC-Kandidaten keine vollständige Abdeckung gewährleisten kann. Die Abschätzung der Anzahl übersehener KRCs ermöglicht mithin die Bewertung des erstellten Goldstandards in Bezug auf seine Abdeckung und erlaubt somit eine Einschätzung, ob der Goldstandard zur Evaluation von Extraktionsexperimenten sinnvoll herangezogen werden kann.

Zur Schätzung des Abdeckungsgrads werden aus den erstellten und in Einzelsätze gesplitteten Subkorpora (s. Abschnitt 4.4) zunächst mit Hilfe eines Perl-Skripts zufällig 500 Sätze ausgewählt („Ziehen mit Zurücklegen“). Jede dieser Stichproben wird alsdann auf das Vorliegen von während der Annotation nicht beachteten KRC-Kandidaten untersucht, wobei in der Annotation berücksichtigte KRCs als erwartete Positive und Nicht-KRCs als erwartete Negative bewertet werden. In die

Auswertung gehen also nur solche Sätze ein, die möglicherweise KRCs darstellen, während des Annotationsprozesses aber nicht berücksichtigt wurden. Zur Abschätzung des Anteils an nicht identifizierten KRC-Kandidaten wird danach auf Grundlage der Stichprobe ein 95%-Konfidenzintervall für den Anteil nicht erkannter KRC-Kandidaten ermittelt. Mit Hilfe eines solchen Konfidenzintervalls kann eine Obergrenze für den Anteil nicht entdeckter KRC-Kandidaten in den Goldstandard-Korpora geschätzt werden. Die entsprechenden Werte können beispielsweise aus einer Tabelle im Lehrbuch von Sachs und Hedderich (2009[13]: 294) nachgeschlagen werden. Tabelle 30 fasst die Ergebnisse der Analyse für alle vier Subkorpora zusammen.

	Sub-EKR	**Sub-TKR**	**Sub-EKD**	**Sub-TKD**
Anzahl nicht bekannter KRC-Kandidaten	7[66]	7	6[67]	10
Stichproben-Anteil	0,0140	0,0140	0,0120	0,0200
95%-Konfidenz-intervall für den Anteilswert	[0,0056, 0,0288]	[0,0056, 0,0288]	[0,0044, 0,0259]	[0,0096, 0,0365]

Tabelle 30: Unbekannte KRC-Kandidaten anhand von einfachen Zufallsstichproben aus allen vier Subkorpora

Für jede der untersuchten Stichproben liegt der Anteil nicht erfasster KRC-Kandidaten deutlich unter fünf Prozent. Ebenso liegt für jede Stichprobe die Obergrenze des jeweiligen Konfidenzintervalls in einem sehr niedrigen Bereich, so dass nur von geringen Abweichungen zwischen dem geschätzten und dem tatsächlichen Anteil nicht erfasster KRC-Kandidaten im Goldstandard ausgegangen werden muss. Angesichts des für alle untersuchten Stichproben geringen Anteils nicht identifizierter KRC-Kandidaten kann insgesamt davon ausgegangen werden, dass durch die beschriebene Annotationsstrategie die in den Subkorpora vorhandenen KRCs in ausreichendem Maße abgedeckt werden.

4.2.5 Zusammenfassung

Aufgrund des Fehlens einer allgemein anerkannten Methodologie für diese Aufgabe hat sich die Erstellung eines Goldstandards als überaus schwierig erwiesen. Eine zusätzliche Schwierigkeit bei der Annotation der Daten erwuchs zudem aus den Besonderheiten von Internet-Daten, die sich tatsächlich durch Redundanz (Dubletten, sehr ähnliche Sätze) sowie – den einzelnen Verarbeitungsschritten geschuldet – durch fehlerhafte Formatierungen auszeichnen. Nicht zuletzt scheint auch der *crawler* „Babouk“ in verschiedenen Durchgängen Ergebnisse unterschiedlicher Qualität geliefert zu haben. Für den Umgang mit diesen Spezialfällen waren Problemlösungsstrategien zu entwickeln. Weitere Probleme waren die fachlich-inhaltliche Bewertung von KRC-Kandidaten sowie die Durchführung der Annotation selbst.
Für den erstgenannten dieser Aspekte – den Umgang mit Internetdaten – wurde ein möglichst exhaustiver Ansatz gewählt, d. h. es wurden zwar keine KRC-Kandidaten doppelt in den Goldstandard aufgenommen, wohl aber sehr ähnliche Varianten, wobei fehlerhafte Formatierungen – soweit sie das Verständnis nicht beeinträchtigen – als irrelevant beibehalten wurden. Vorstehende Ausführungen dürften zudem deutlich gemacht haben, dass eine fachliche Bewertung der KRC-

[66] Die unbekannten KRC-Kandidaten für die russischen Korpora sind in Anhang 6 zusammengefasst.
[67] Die unbekannten KRC-Kandidaten für die deutschen Korpora sind in Anhang 7 zusammengefasst.

Kandidaten nicht durchgeführt wurde. Auch dies entspricht allerdings einem übersetzungsbezogenen Szenario, in dem Übersetzer Quellen terminologisch relevanter Informationen recherchieren und dann – in der Regel und sofern keine zuverlässigeren Informationsquellen zur Verfügung stehen – auf Grundlage ihrer Suchergebnisse einen Begriff erlernen und ggf. in ihrer Datenbank ablegen. Für diese Studie bedeutet dies konkret, dass die Schritte Korrektheits- und Plausibilitätsprüfung *nach* der KRC-Extraktion durchgeführt werden müssen.

Der Status der erarbeiteten Annotation bzw. vielmehr der Bewertung der von Annotator A ausgewählten KRC-Kandidaten durch vier weitere Annotatoren pro Sprache war bereits kurz diskutiert worden. Im statistischen Sinne verallgemeinerbare Ergebnisse konnten dabei aufgrund eingeschränkter Ressourcen nicht erarbeitet werden, so dass der letztlich erstellte Goldstandard eher als Grundlage für spätere Studien zu betrachten ist, in denen nicht nur stringentere Annotationsrichtlinien verwendet, sondern auch die Ergebnisse der in den folgenden Kapiteln noch zu beschreibenden Analysen überprüft – belegt oder widerlegt – werden können. Die hier durchgeführte Annotationsstudie scheint zumindest frühere Ausführungen – etwa die Cramers (2011) – dahingehend zu bestätigen, als bei der Bewältigung der sehr komplexen Annotationsaufgabe nicht nur linguistische Faktoren zu berücksichtigen sind. Unter „stringenteren“ Annotationsrichtlinien werden folglich insbesondere Strategien für den Umgang mit Zweifelsfällen – etwa im Fall unbekannter Fachgebiete – verstanden, weniger detaillierte linguistische Beschreibungen. Insofern, als die Beschreibung und Extraktion wissenshaltiger Kontexte hier insbesondere im Zusammenhang mit Fragen der Benutzbarkeit und Nutzungsqualität terminologischer Datensammlungen sowie der Terminologierecherche betrachtet wird, ist zudem auf die Bedeutung von Nutzerstudien hinzuweisen, deren Ergebnisse auch für die hier untersuchte Problemstellung wichtige Hinweise liefern könnten.

Im Ergebnis des, wie ausgeführt, iterativen Annotationsprozesses wurde ein Goldstandard auf Grundlage von inhaltlich diversifizierten Internet-Korpora mittleren bis populären Fachlichkeitsgrads (s. hierzu auch die Ausführungen im nachfolgenden Abschnitt) erstellt. Die Bewertungen der Annotatoren können vor diesem Hintergrund als intersubjektives Votum betrachtet werden, das zumindest einen auf wenige Personen beschränkten Konsens oder Dissens abbildet. Vorsichtig – die Formulierung inhaltlicher und methodischer Vorbehalte bleibt dennoch angebracht – kann festgestellt werden, dass das spezifische, definitionsbezogene Textsortenwissen, von dem Cramer (ebd.) spricht, tatsächlich auch im Kreis der Studienteilnehmer in Grundzügen zu existieren scheint. Des Weiteren zeigen die Ergebnisse in Tabelle 29, insbesondere die umfänglichen Bereinigungen im Sinne des Entfernens „falscher Positive“, dass Annotator A selbst das eigene Verständnis für die Kategorie des wissenshaltigen Kontexts erst im Zuge der Durchführung der Annotation selbst schärfen konnte, was wiederum für die Schwierigkeit der Beschreibung dieser Kategorie spricht. Allerdings scheint Tabelle 30 auf eine erfreulich gute Abdeckung der KRCs in den Subkorpora hinzudeuten.

4.3 Analyse des Goldstandards

4.3.1 Theoriegeleitete Klassifikation von Textsorten in Web-Korpora

Im weiteren Verlauf vorliegender Untersuchung sollen wissenshaltige Kontexte nur aus jenen Subkorpora untersucht werden, in denen alle Sätze des Goldstandards enthalten sind. Diese

Subkorpora wurden, wie in Abschnitt 4.2.1 beschrieben, nach der Auswertung des ersten Annotationsdurchgangserstellt. In diesem Zusammenhang erscheint es angemessen, auch die Subkorpora genauer zu untersuchen. Für die Dokumentation der Subkorpora sollen insbesondere neben numerischen Daten auch Informationen über die in den Subkorpora enthaltenen Texttypen und Textsorten gesammelt werden, um so die bereits erwähnten Hypothesen über das Auftreten wissenshaltiger Kontexte genauer zu beleuchten. Dabei wird zunächst mit Nord (1995[3]: 80) unter „Texttyp“ eine Klasse von Texten mit einer gemeinsamen, sie dominierenden Sprachfunktion verstanden, wohingegen Textsorten „[...] einen kommunikativen Handlungstyp realisieren [...]“ und entsprechende konventionalisierte sprachlich-textuelle Merkmale aufweisen. Für eine Charakterisierung der in den Korpora enthaltenen Texte lassen sich die in Tabelle 5 dargestellten Kriterien zur Beschreibung von Korpustexten nach Atkins et al. (1992) zu etablierten Fachtexttypologien wie etwa derjenigen von Göpferich (1995: 123 ff.) in Beziehung setzen und durch im Internet häufig vorkommende Textsorten ergänzen. Tabelle 31 gibt einen Überblick über die so entstehende Typologie von Texten in spezialisierten *web*-Korpora und nennt gleichzeitig typische in den vier Korpora aufgefundene Textsorten. Durch Kursivschrift hervorgehoben sind jeweils grundlegende Texttypen, die alsdann – Göpferichs Typologie folgend – in Untertypen aufgegliedert werden.

Texttyp	**Textqualität**	**Kommunikationssituation**	**Textsorte**
Fortschrittsorientiert-aktualisierend			
Publizistisch aufbereitet	Mittel	Freizeit, Arbeitsplatz	Artikel
	Hoch	Freizeit, Arbeitsplatz	Firmen- oder Produktvorstellung
faktenorientiert	Hoch	Arbeitsplatz	Projektentwurf
	Hoch	Arbeitsplatz, Ausbildung	Wissenschaftlicher Beitrag
Instruktiv[68]			
Theoretisches Wissen vermittelnd: Das Interesse weckend	Mittel	Ausbildung	Aufgabe
Theoretisches Wissen vermittelnd[69]	Mittel	Freizeit, Arbeitsplatz, Ausbildung	Erklärender Text
Mensch/Technik-interaktions-orientiert	Mittel	Freizeit, Arbeitsplatz	Betriebsanleitung
	Mittel	Freizeit, Arbeitsplatz	Do it yourself
	Niedrig	Freizeit, Arbeitsplatz, Ausbildung	Forum
Juristisch-normativ			
	Hoch	Arbeitsplatz, Ausbildung	Normen, Standards, Rechtsvorschriften
	Hoch	Arbeitsplatz	Patent

[68] Göpferich verwendet für diesen Texttyp den Terminus „didaktisch-instruktiv“. Die Gründe für die terminologische Modifikation werden weiter unten erläutert.

[69] Auf die von Göpferich verwendete Charakterisierung „mnemotechnisch organisiert“ wird hier aus Gründen verzichtet, die ebenfalls weiter unten erläutert werden sollen.

Wissenszusammen-stellend			
Satzfragmentarisch	Hoch	Freizeit, Arbeitsplatz, Ausbildung	Wörterbuch
	Mittel	Nicht eindeutig zuordenbar	Kurztext
	Mittel	Nicht eindeutig zuordenbar	Liste
Enzyklopädisch	Mittel	Freizeit, Ausbildung	Biographie

Tabelle 31: Klassifikation der in den Web-Korpora vorgefunden Texte

Auf den ersten Blick fällt auf, dass die in den Korpora aufgefunden Texte nicht alle der von Göpferich beschriebenen, im Gegenzug allerdings einige internetspezifische Textsorten abdecken. Allen in Tabelle 31 angeführten Textsorten gemeinsam sind darüber hinaus die folgenden Eigenschaften:

- Modalität: schriftlich,
- Herkunft: häufig nicht feststellbar,
- Faktizität: an reale Sachverhalte gebunden,
- Spezialisierungsgrad (im Sinne von „Babouk“): hoch,
- Veröffentlichungszeitpunkt: häufig nicht eindeutig feststellbar.

Die Zuordnung der Textsorten zu den Texttypen folgt weitgehend Göpferichs Typologie, enthält allerdings auch einige Besonderheiten, etwa die Textsorte Aufgabe, wie sie beispielsweise in *eLearning*-Inhalten zu finden ist: Aufgabentexte erläutern Zusammenhänge und fragen gleichzeitig spezifisches Wissen ab, weisen also auch dialogisch-didaktische Elemente auf. Der Spezifik des Mediums Internet geschuldet ist zudem der Umstand, dass die Korpora neben klassischen Lehrbüchern (etwa aus PDF-Quellen) auch eine Reihe anderer gezielt Wissen darstellender Texte enthalten, die nicht dem Lehrbuchtyp entsprechen, beispielsweise Diplomarbeiten oder Beschreibungen von Schulprojekten, aber auch recht kurze Texte mit informativen Merkmalen. Diese Texte wurden gemeinsam mit klassischen Lehrtexten in der Textsorte erklärende Texte zusammengefasst, wobei allerdings auf das von Göpferich verwendete Attribut der mnemotechnischen Organisation verzichtet werden musste, um diese sprachlich wie formal recht disparaten, im Hinblick auf die dominierende Textfunktion – spezifische Informationen zu vermitteln – jedoch homogenen Texte in einer gemeinsamen Kategorie unterzubringen. Auch von einem didaktischen Zusammenhang im Sinne einer institutionell formalisierten Ausbildungssituation, in deren Rahmen Wissen vermittelt wird, kann hier nicht immer die Rede sein. Deutlich ist dennoch das Anliegen, gesammeltes Fachwissen verständlich darzustellen. Daher wurde dieser Texttyp in Abweichung von Göpferichs Typologie als „instruktiv“ bezeichnet.

Betriebsanleitungen bilden eine Mischform aus erklärenden und anleitenden Elementen, allerdings ist hier anzumerken, dass im Zusammenhang mit Internettexten die für Betriebsanleitungen normalerweise typische Abgeschlossenheit (Broschürenform) und Organisation nicht zu prüfen ist, so dass auch relativ kurze Texte anleitenden Charakters als Betriebsanleitungen klassifiziert wurden. „*Do it yourself*“-Texte haben dagegen einen rein anleitenden Charakter, der sich beispielsweise für das Russische in der häufigen Verwendung von Imperativformen widerspiegelt. Forumstexte sind meist dialogisch oder polyphon (d. h. durch aufeinander bezogene Äußerungen einer ganzen Autorengruppe gekennzeichnet) und oft von niedriger sprachlicher Qualität, behandeln allerdings – etwa im IT-Bereich – häufig konkrete technische Probleme, so dass sie dem Typus der interaktionsorientierten Texte zugeordnet wurden. Als Kurztexte wurden sehr kurze bzw.

unvollständige Textelemente eingestuft. Listen sind enumerativer Natur und oft ebenfalls fragmentarisch, allerdings sind sie eine für das Internet typische Präsentationsform, da sie – etwa auf den Startseiten von *homepages* oder Portalen – einen Überblick über verschiedene Themen oder Nachrichten geben. Unter diesem Gesichtspunkt können in seltenen Fällen auch als Listen klassifizierte Texte relevante Informationen enthalten. Für die hier interessierende Aufgabe – die Beschreibung und Extraktion wissenshaltiger Kontexte – sind jedoch im Einklang mit ihrer dominanten Textfunktion insbesondere wissenschaftliche Beiträge, erklärende Texte und Betriebsanleitungen sowie Patente, ferner auch Aufgaben und eventuell Artikel von Interesse. Generell scheinen die Bemerkungen Kalgrens (2010: 39 f.) zur Spezifik von Web-Texten auch auf die hier gesammelten Daten zuzutreffen, wodurch die Klassifikation der aufgefundenen Textsorten nach einem einheitlichen Schema erschwert wird: Die tendenzielle Vermischung kommerzieller und nicht-kommerzieller Inhalte und die Vielzahl von interaktiven Kommunikationsformen bis hin zu *eLearning*-Angeboten (etwa die Textsorte „Aufgabe“ in Tabelle 31) führen zur Entstehung von Textsorten, die sich durch Textsortentypologien für Printmedien kaum erfassen lassen. Aus terminographischer Sicht nachteilig erscheint der Umstand, dass der zeitliche und kommunikative Kontext der Produktion der durch „Babouk“ gesammelten Texte nur in Ausnahmefällen nachvollziehbar ist. Hieraus ergeben sich methodologische Unschärfen, die im Zusammenhang mit einem *„web as corpus“*-Ansatz nicht aus dem Blick verloren werden dürfen, sondern stets mitgedacht werden müssen: Im Sinne des bereits diskutierten korpuslinguistischen Ansatzes sind Bewertungen von Inhalten weniger an extralinguistische als vielmehr an quantitative Kriterien zu knüpfen. Für eine solche Vorgehensweise bietet der hier beschriebene Goldstandard jedoch aufgrund seiner geringen Größe nur eingeschränkt Ansatzpunkte.

Im Zusammenhang mit der Beschreibung von Textsorten in den gesammelten Korpora erscheinen noch zwei Anmerkungen von Bedeutung: Zwar wurde allen Textsorten der Spezialisierungsgrad „hoch“ zugeordnet, allerdings wurde dabei keines der üblicherweise verwendeten Kriterien zur Beschreibung von Fachtexten (vgl. Roelcke 2005[2]) zugrunde gelegt, da sie nicht den Selektionskriterien des verwendeten *crawlers* „Babouk“ entsprechen: „Babouk“ selektiert Texte mit hoher Terminus-Dichte und es ist dieser Umstand, der die hier beschriebenen Textsorten im Sinne des Ansatzes dieser Arbeit zu Fachtexten macht. Offensichtliche Binnendifferenzierungen der Fachlichkeit – etwa zwischen *„do-it-yourself“*-Texten und wissenschaftlichen Arbeiten – werden dabei bewusst nicht berücksichtigt, da sie von „Babouk“ nicht erkannt werden.

Darüber hinaus handelt es sich bei der hier vorgeschlagenen Textsortenklassifikation nicht um eine abschließende Beschreibung der in *web*-Korpora auftretenden Texte, sondern vielmehr um eine theoriegeleitete *ad-hoc*-Klassifikation, welche die Beschreibung und – im Sinne des oben zitierten Ansatzes von Ahmad und Rogers (2001) – die Kritik der erstellten Korpora ermöglichen soll. Die Erarbeitung einer inhaltlich abgeschlossenen Textsortentypologie für spezialisierte Web-Korpora würde nicht nur den Rahmen vorliegender Arbeit sprengen, sondern scheint auch aus methodisch-theoretischer Sicht kaum realisierbar. Beispielweise unterstreicht Roelcke (2005[2]: 47 ff.) die Relativität und inhaltliche Bindung jedes textsortentypologischen Ansatzes an eine konkrete wissenschaftliche Zielstellung:

> Es wird deutlich, „[...] dass die zahlreichen Fachtextsorten, die sich im Laufe der Fach- und Sprachgeschichte herausgebildet haben, nur schwer durch eine wie auch immer angelegte Systematik erfasst werden können, da diese kaum eine hinreichende Differenzierung dieser Textsortenvielfalt zulässt. Vor diesem Hintergrund muss eine fachsprachenlinguistische Textsortengliederung zu dem gegenwärtigen Stand der Forschung immer an den wissenschaftlichen Fragestellungen ausgerichtet

sein, unter denen sie jeweils zum Einsatz kommt; eine Fachtextsortengliederung per se erscheint also kaum sinnvoll."

Im Sinne Roelckes ist die vorgeschlagene Klassifikation angemessen, da sie eine Abschätzung der Korpusqualität im Hinblick auf das Vorhaben der Extraktion wissenshaltiger Kontexte ebenso ermöglicht wie die Untersuchung der Verteilung wissenshaltiger Kontexte in Korpora. Pearson (1998) und Meyer (2001) zufolge ist mit dem Auftreten relevanter Kontexte insbesondere in den Textsorten Betriebsanleitung und erklärender Text sowie eingeschränkt in wissenschaftlichen Beiträgen zu rechnen. Korpora mit einem hohen Anteil dieser Textsorten sind daher u. U. durch eine höhere Kontextdichte gekennzeichnet.

4.3.2 Eigenschaften der Subkorpora

Tabelle 32 gibt nun einen Überblick über die Anzahl der Texte, welche die entsprechenden Subkorpora bilden, sowie über deren Größe, gemessen als Anzahl der laufenden Wörter in den Subkorpora. Auch der Anteil von Wikipedia-Texten an der Gesamtzahl der Texte ist für jedes Subkorpus ausgewiesen. Darüber hinaus enthält die Tabelle auch Informationen über die Anzahl der Sätze in den Subkorpora sowie die durchschnittliche Satzlänge. Diese Werte wurden auf dieselbe Art und Weise ermittelt wie in Abschnitt 4.1.4.1, wobei die Anzahl der Wörter auf den bereits mit Lingua::Sentence bearbeiteten Korpora ermittelt wurde.

Subkorpus	Texte	Wikipedia	Textsorten: Texte	Wörter[70]	Sätze	Satzlänge
Sub-EKR	36	1	• **Betriebsanleitung: 16** • **Erklärender Text: 16** • Artikel: 1 • Firmen- oder Produkt-vorstellung: 1 • Liste: 1 • Wörterbuch: 1	98.527	8860	11,12
Sub-TKR	36	6	• **Erklärender Text: 26** • Firmen- oder Produktvorstellung: 4 • Standard: 2 • Betriebsanleitung: 1 • Liste: 1 • Projektentwurf: 1 • Wissenschaftlicher Beitrag: 1	75.053	6640	11,30
Sub-EKD	57	1	• **Erklärender Text: 34** • Firmen- oder Produkt-vorstellung: 8 • Liste: 7 • Artikel: 3 • Wissenschaftlicher Beitrag: 2 • Kurztext: 1 • Patent: 1 • Projektentwurf: 1	159.937	9980	16,03

[70] Auch diese Werte wurden mit Hilfe der Linux-Funktion wc ermittelt.

Sub-TKD	24	19	• **Erklärender Text: 22** • Aufgabe: 1 • Firmen- oder Produktvorstellung: 1	172.163	12.344	13,95

Tabelle 32: Überblick über Subkorpora des Goldstandards

Tabelle 32 gibt Hinweise auf die Struktur der für den Goldstandard erstellten Subkorpora. Freilich handelt es sich bei der hier dargestellten Bestimmung von Textsorten lediglich um eine Annäherung, die aus Gründen der Praktikalität auf eine intersubjektive und damit möglicherweise besser nachvollziehbare Vorgehensweise verzichten musste, allerdings im Hinblick auf die Charakterisierung der erstellten Sub-Korpora dennoch brauchbare Hinweise liefert. Bei dem zu untersuchenden Subkorpus des EKR handelt es sich wie bereits im Falle des Gesamtkorpus‘ um ein inhaltlich konzentriertes Korpus mit fachlich-erklärenden Texten. Das Sub-TKR ist ein sehr kleines Korpus mit weit überwiegend erklärenden Texten. Bei der Markierung wissenshaltiger Kontexte im Sub-EKD ergab sich in Übereinstimmung mit den dargestellten Hypothesen zur Verteilung wissenshaltiger Kontexte allerdings ein Schwerpunkt auf erklärenden Texten, der nicht der Textsortenverteilung des EKD insgesamt zu entsprechen scheint, angesichts der Gesamtstruktur des EKD enthält jedoch auch das entsprechende Subkorpus eine hohe Anzahl vergleichsweise kurzer Texte. Für das Sub-TKD ist insbesondere der überdurchschnittlich große Anteil an Wikipedia-Texten kennzeichnend, die jeweils eine große Anzahl von KRCs zu enthalten scheinen. Dies entspricht einer der in Tabelle 6 dargestellten Eigenschaften des Gesamtkorpus‘. Der hohe Anteil von Wikipedia-Texten legt im Zusammenhang mit der Tatsache, dass das TKD-Subkorpus im Vergleich zu den anderen Subkorpora deutlich weniger Einzeltexte aufweist, nahe, dass in Wikipedia-Texten vergleichsweise mehr wissenshaltige Kontexte aufgefunden werden können als in anderen *web*-Texten.
Im Hinblick auf die Verteilung der Textsorten, in denen wissenshaltige Kontexte aufgefunden wurden, scheinen die Subkorpora sowohl Eigenschaften des Internet selbst als auch des untersuchten Fachgebiets zu spiegeln. So scheinen für das russische Internet im Bereich KFZ insbesondere Betriebsanleitungen und sonstige Instruktionen eine hervorragende Bedeutung einzunehmen. Für das EKD kann dagegen – angesichts der lebhaften öffentlichen Diskussion um erneuerbare Energien keineswegs überraschend – ein nachhaltiges kommerzielles sowie wissenschaftliches Interesse an Windkraft festgestellt werden, das sich in einer größeren Zahl an Firmen- und Produktvorstellungen, wissenschaftlichen Texten und Patenten niederschlägt. Insgesamt scheinen die hier ausgewählten Texte die Hypothesen Pearsons und Meyers zum Auftreten wissenshaltiger Kontexte tendenziell zu bestätigen. Festgehalten werden kann zwar, dass KRCs auch – mitunter überraschend – in anderen Textsorten auftreten können, das Gros aber tatsächlich in instruktiven Texten zu finden ist. Im Zusammenhang mit den ausgewählten Texten und den markierten Kontexten ist noch eine Anmerkung anzubringen: Das EKR-Subkorpus enthält auch einen als „Wörterbuch“ klassifizierten Text. Dieser wiederum enthält einen hohen Anteil an Begriffsdefinitionen der in 15) dargestellten Form:

15) Вал. Деталь машины, вращающаяся в подшипниках, служащая опорой и передающая крутящий момент.
[Ü*bersetzung:* Welle. Sich in Lagern drehendes Maschinenteil, das als Stütze sowie zur Übertragung des Drehmoments dient.]

Es handelt sich hierbei um eine Definition in klassischer Wörterbuchform, in welcher das Definiendum als Lemma der eigentlichen Definition vorangestellt und die Definition in sprachlich verknappter Form formuliert ist. Solche Definitionen waren in Kapitel 2 aus der Extension des Begriffs „wissenshaltiger Kontext“ ausgeschlossen worden, da Wörterbuchdefinitionen aufgrund ihrer spezifischen Form aus Extraktionssicht an andere Anforderungen geknüpft sind als wissenshaltige Kontexte und hinsichtlich ihres Informationsgehalts anderen Anforderungen unterliegen. Nun enthält derselbe Text aber auch KRCs, wie sie auch außerhalb von Wörterbüchern durchaus auftreten können, ein Beispiel dafür ist 16:

16) Выносливость — способность материалов и конструкции сопротивляться разрушению при действии повторно переменных напряжений.
[*Übersetzung:* Die Ermüdungsfestigkeit ist die Fähigkeit von Werkstoffen oder Konstruktionen, unter Einwirkung wiederholt wechselnder Einflüsse (*wörtlich*: Spannungen) einer Zerstörung standzuhalten.]

Ein solcher KRC kann durchaus auch außerhalb von Wörterbüchern auftreten und wurde aus diesem Grund auch in den Goldstandard aufgenommen. Da der fragliche Text mithin eine Mischung aus Wörterbuchdefinitionen und validen KRCs enthält, wurde er in das EKR-Subkorpus aufgenommen. Tabelle 33 stellt die Verteilung der (nur grob bestimmten) Fachgebiete in den Korpora dar. Listen der URLs für die Texte in den russischen und deutschen Subkorpora finden sich in Anhang 8 bzw. 9.

Subkorpus	**Fachgebietsverteilung**
Sub-EKR	weitgehend monothematisch (KFZ)
Sub-TKR	• Ingenieurwesen: 187 Kontexte • Atomenergie: 20 Kontexte • Energieversorgung: 19 Kontexte • Alternative Energie: 18 Kontexte • Physik und Chemie: 10 Kontexte • Raumfahrt: 10 Kontexte • Windenergie: 3 Kontexte • Militärtechnik: 1 Kontext
Sub-EKD	weitgehend monothematisch (Windenergie)
Sub-TKD	• Ingenieurwesen: 121 Kontexte • Physik und Chemie: 54 Kontexte • Alternative Energie: 44 Kontexte • Militärtechnik: 22 Kontexte • Energieversorgung: 20 Kontexte • Normierung: 13 Kontexte • Recht: 11 Kontexte • Atomenergie: 7 Kontexte • Management: 2 Kontexte • Geographie: 1 Kontext

Tabelle 33: Fachgebietsverteilung in den Subkorpora des Goldstandards

4.4 Vorverarbeitung, Annotation und Formatierung der Korpora

Die Beschreibung und Extraktion wissenshaltiger Kontexte erfordert Textkorpora, die annotiert, also mit expliziten Deskriptionen linguistischer Strukturen und dem Text anhaftender linguistischer Wissensbestände versehen wurden. Neben dieser Anreicherung der Korpora mit linguistischer Information, welche die zielgerichtete Suche nach sprachlichen Strukturen und Mustern ermöglicht, ist es im Umgang mit *web*-Korpora – wie bereits dargestellt – im Sinne der Analysequalität empfehlenswert, Satzduplikate weitgehend zu entfernen, bevor mit der eigentlichen Analyse begonnen wird. Der linguistischen Annotation der Korpora sollten daher Vorverarbeitungsschritte vorausgehen, in deren Verlauf die Korpora auf die weitere Analyse vorbereitet werden (*preprocessing*). In einem letzten Verarbeitungsschritt, dem *postprocessing* (vgl. Baker et al. 2006: 132), schließlich sind die während der Annotation der Korpora erhaltenen Informationen in ein Format zu bringen, das eine leichte Weiterverarbeitung der Daten im Sinne der hier interessierenden Fragestellung erlaubt. Zur linguistischen Aufbereitung der russischen und deutschen Korpora wird in diesem Kapitel ein Satz existierender, frei verfügbarer Werkzeuge genutzt. Als Programmiersprache für eigene Skripte wird zudem durchgängig Perl verwendet. Abbildung 9 stellt die Bearbeitungsschritte graphisch dar. Im Folgenden werden die einzelnen Schritte eingehender erläutert.

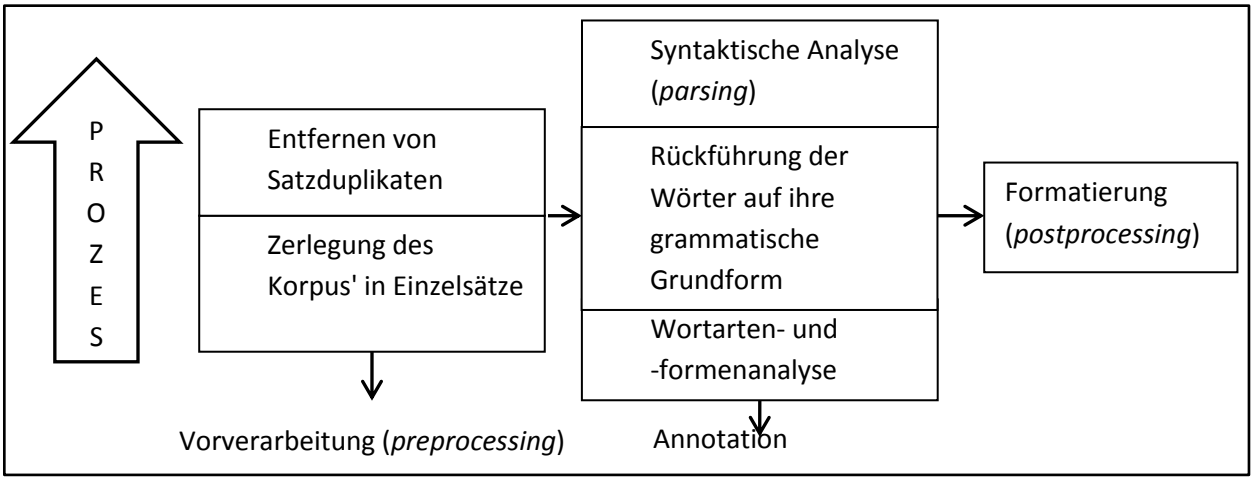


Abbildung 9: Vorverarbeitung, linguistische Annotation und Formatierung der Korpora

4.4.1 Vorverarbeitung

Die Zerlegung eines Korpus' in Einzelsätze ist eine wesentliche Voraussetzung für die Analyse sprachlicher Regularitäten auf Satzebene. Für die Erkennung der Satzgrenzen in russischen und deutschen Textkorpora wird in vorliegender Arbeit das Perl-Modul Lingua::Sentence[71] genutzt. Es handelt sich hierbei um ein einfach zu implementierendes Werkzeug, das Satzgrenzen anhand linguistischer Regeln erkennt und den Text durch Einfügen von Zeilenumbrüchen in Einzelsätze auftrennt. Sogenannte *nonbreaking prefixes* dienen dabei zur Formulierung von Ausnahmeregeln für Fälle, in denen kein Zeilenumbruch einzufügen ist (etwa bei Abkürzungen).

[71] http://search.cpan.org/~achimru/Lingua-Sentence-1.05/lib/Lingua/Sentence.pm, Zugriff am 24.03.2014.

Das Modul bietet bereits Funktionalitäten für die Bearbeitung deutscher Korpora, für die Analyse russischer Texte war zum Zeitpunkt der Korpusannotation allerdings eine neue Datei mit *nonbreaking prefixes* zu erstellen. Einige linguistische Regeln konnten hierfür dem frei verfügbaren Programm „Language Tool“[72], das Werkzeuge für die Stil-, Grammatik- und Rechtschreibprüfung für eine Reihe unterschiedlicher Sprachen anbietet, entnommen werden, weitere 124 Abkürzungen wurden aus dem multilingualen Online-Wörterbuch Wiktionary[73] übernommen.

Eine besondere Herausforderung für die korrekte Zerlegung der gesammelten Korpora in Einzelsätze ist der Umstand, dass aus dem Internet heruntergeladene Texte – möglicherweise infolge von Formatkonvertierungen oder von den Autoren eingefügte *hard returns* – häufig falsche Zeilenumbrüche enthalten. Darüber hinaus sind für die Beschreibung und Extraktion wissenshaltiger Kontexte u. U. auch Listen von Interesse, beispielsweise dann, wenn eine solche Liste eine Aufzählung relevanter Eigenschaften (Attribute) eines Begriffs darstellt. Bei der Annotation des Goldstandards wurden daher auch solche Kontextkandidaten als valide angesehen, die einer dieser beiden Gruppen zuzuordnen sind, also inmitten des eigentlichen Satzes Zeilenumbrüche enthalten, sonst aber die erarbeiteten Validitätskriterien für wissenshaltige Kontexte erfüllen. Um im Rahmen eines Verfahrens für die Extraktion wissenshaltiger Kontexte solche KRCs möglichst einfach, d. h. ohne Rekurs auf über die Satz- bzw. die Zeilenebene hinausgehende linguistische Analyseebenen, identifizieren zu können, wurden die Perl-Skripte für die Zerlegung russischer und deutscher Korpora in Einzelsätze daraufhin optimiert, solche typographisch problematischen Sätze möglichst als Einzelsatz zu segmentieren, also fehlerhafte Zeilenumbrüche oder Umbrüche in Listen zu entfernen, ohne dabei korrekte Segmentierungen zu verlieren. Der Vorteil des verwendeten Perl-Moduls besteht nun darin, dass es die für die Bearbeitung russischer Texte erforderliche Integration zusätzlicher linguistischer Regeln ohne größere Probleme ermöglicht sowie im Sinne einer zielgerichteten Optimierung der Satzsegmentierung in Bezug auf den verwendeten Regelsatz leicht modifizierbar ist. Insgesamt wurden 243 russische *nonbreaking prefixes* sowie eine Reihe zusätzlicher Trennungsregeln und Regeln für die nachträgliche Korrektur fehlerhafter Zeilenumbrüche definiert. Auch für das Deutsche wurden bei der Implementierung des Moduls zusätzliche Trennungs- bzw. Korrekturregeln festgelegt.

Da es sich bei der Satzsegmentierung der Korpora insofern um einen wesentlichen Vorverarbeitungsschritt handelt, als dessen Ergebnis maßgeblich für den Erfolg späterer Annotationsschritte ist – beispielsweise können im Falle eines Fehlens erforderlicher Satztrennungen syntaktische Dependenzstrukturen nicht mehr mit ausreichender Qualität ermittelt werden –, und da das Perl-Modul für die Bearbeitung deutscher und russischer Korpora um einige Regeln ergänzt wurde, soll das Resultat der beschriebenen Optimierung hier kurz evaluiert werden. Für die Evaluation wurden die im vorigen Kapitel erstellten Sub-Korpora in Einzelsätze zerlegt, und zwar in insgesamt drei Szenarien:

- ohne jegliche zusätzliche Regeln,

[72] http://www.languagetool.org/. Mit Hilfe von Achim Ruopp konnten die Informationen aus einer Datei mit Segmentierungsregeln für Language Tool extrahiert werden. Eine aktuelle multilinguale Segmentierungsdatei findet sich hier: https://github.com/languagetool-org/languagetool/blob/master/languagetool-core/src/main/resources/org/languagetool/resource/segment.srx, Zugriff am 24.03.2014.

[73] http://en.wiktionary.org/w/index.php?title=Category:Russian_abbreviations&from=%D0%90, Zugriff am 24.03.2014.

- nur mit einer zusätzlichen Liste von Abkürzungen, nach denen kein Zeilenumbruch eingefügt werden soll, sowie einer Normalisierung der Interpunktion (d. h. dem Entfernen von Leerzeichen zwischen Wörtern und Abkürzungspunkten), damit die Abkürzungen auch erkannt werden,
- mit dem vollen Satz der oben beschriebenen Korrektur- und Trennungsregeln.

Anhand von 100 KRCs aus dem jeweiligen Goldstandard wurde die Satzsegmentierung alsdann für jedes Sub-Korpus evaluiert. Tabelle 34 fasst die Ergebnisse zusammen.

	Sub-EKD	**P**	**R**	**Sub-TKD**	**P**	**R**	**Sub-EKR**	**P**	**R**	**Sub-TKR**	**P**	**R**
Zerlegung ohne Zusatzregeln												
Korrekt segmentierte Sätze	87			70			92			92		
Falsche Zeilenumbrüche	5			27			4			6		
Fehlende Zeilenumbrüche	8	**0,95**	**0,87**	3	**0,73**	**0,70**	4	**0,96**	**0,92**	2	**0,94**	**0,92**
Zerlegung mit Normalisierung und Abkürzungsliste												
Korrekt segmentierte Sätze	86			72			95			91		
Falsche Zeilenumbrüche	5			22			0			5		
Fehlende Zeilenumbrüche	9	**0,94**	**0,85**	6	**0,78**	**0,72**	5	**1**	**0,95**	4	**0,95**	**0,91**
Zerlegung mit zusätzlichen Trennungs- und Korrekturregeln												
Korrekt segmentierte Sätze	85			81			98			98		
Falsche Zeilenumbrüche	0			4			0			1		
Fehlende Zeilenumbrüche	15	**1**	**0,85**	15	**0,96**	**0,81**	2	**1**	**0,98**	1	**0,99**	**0,98**

Tabelle 34: Evaluation für die Satzsegmentierung der russischen und deutschen Sub-Korpora

Die Tabelle stellt für jedes der drei Segmentierungsszenarien dar, welchen Anteil jeweils die korrekt segmentierten Sätze sowie die fälschlicherweise nicht eingefügten oder aber an falscher Stelle eingefügten Zeilenumbrüche ausmachen. Die Spalten *P* und *R* geben darüber hinaus für jedes Szenario und jedes Sub-Korpus den Wert der bereits eingeführten Evaluationsmaße Präzision und Abdeckung (*recall*) an. Für die Berechnung der Präzision der Segmentierung wurde dabei die folgende Formel verwendet:

17) $$P = \frac{Korrekt\ segmentierte\ Sätze + Sätze\ mit\ fehlenden\ Zeilenumbrüchen}{100}$$

Sätze mit fehlenden Zeilenumbrüchen gehen in die Berechnung der Präzision ein, da sie mindestens an einem Ende korrekt gesplittet sind, also einen korrekten Umbruch aufweisen. Der *recall*, also der Anteil der korrekt gesetzten Zeilenumbrüche, entspricht dagegen dem Anteil der vollständig korrekt segmentierten Sätze. Tabelle 34 zeigt, dass für dasjenige Segmentierungsszenario, in dem neben dem Perl-Modul selbst und dessen *non-breaking*-Regeln der volle Satz eigener Trennungs- und Korrekturregeln zum Einsatz kam, stets die höchste Präzision und in den meisten Fällen auch der höchste *recall* erreicht werden konnte.
Bereits im vorigen Kapitel war zudem die Notwendigkeit hervorgehoben worden, Satzduplikate aus *web*-Korpora zu entfernen. Nach der Segmentierung der Korpora in Einzelsätze wird dies durch ein einfaches Perl-Skript geleistet, das mit Hilfe einer Hash-Tabelle 100-prozentige Duplikate aus den Korpora herausfiltert. Die komplexere Filterung sehr ähnlicher Sätze und sogenannter *boilerplate*-Elemente (z. B. Online-Werbung, *footer*, *header*) unterbleibt jedoch, da zwar Werkzeuge für diese Aufgaben existieren, jedoch bestimmte Eingabeformate erwarten, die mit dem von „Babouk“ gelieferten Format nicht kompatibel sind. Es wird erwartet, dass diese Elemente bei den durchzuführenden Analyse- und Extraktionsaufgaben nur von untergeordneter Bedeutung sind. In den Goldstandard wurden zwar aus Gründen der Vollständigkeit auch sehr ähnliche KRCs aufgenommen, die linguistische Analyse sollte aber aufgrund der geringen Anteile dieser Sätze nicht verzerrt werden.

4.4.2 Linguistische Annotation

Zur linguistischen Annotation zählen die Wortartenanalyse, bei der jedes Wort einer Wortartenkategorie zugeordnet wird (Wortarten-*tagging*), sowie die morphologische Wortformenanalyse, bei der die morphologischen Informationen ausgewertet und in Form eines komplexen *tags* annotiert werden. Darüber hinaus zählen auch die Lemmatisierung, bei der jedes Wort auf seine grammatische Grundform zurückgeführt wird, sowie die syntaktische Analyse (*parsing*), in der – abhängig von der Charakteristik des verwendeten Parsers – syntaktische Satzkonstituenten und/oder syntaktische Abhängigkeiten markiert werden, zur linguistischen Annotation.

4.4.2.1 Wortartentagging und Lemmatisierung

Zur Annotation der Wortarten in russischen Korpora wird der bereits erwähnte probabilistische TreeTagger[74] (Schmid 1994) verwendet. Auf Grundlage der Trainingsdaten – einem annotierten Textkorpus – ermittelt dieser *tagger* die Wahrscheinlichkeit eines bestimmten Wortarten-*tags* mit Hilfe eines binären Entscheidungsbaums (ebd.), wobei die zwei zuletzt getaggten Einheiten für die Entscheidung berücksichtigt werden. Ein Modell[75] für die Annotation russischer Korpora mit Hilfe des TreeTaggers wurde – auch dies wurde bereits erwähnt – von Sharoff trainiert. Das dabei verwendete *tagset* (Sharoff et al. 2008) erlaubt die feinkörnige Analyse russischer Wortformen und wurde in

[74] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, Zugriff am 24.03.2014.
[75] http://corpus.leeds.ac.uk/mocky/russian.par.gz/, Zugriff am 24.03.2014.

Übereinstimmung mit den Multext-East-Richtlinien[76] für die Erstellung lexikalischer Ressourcen und annotierter Textkorpora in mittel- und osteuropäischen Sprachen (vgl. Erjavec 2010) entwickelt[77]. Aus diesen feinkörnigen Annotationen können in einem Abstraktionsschritt auch die bloßen Wortarten-*tags* abgeleitet werden. Der TreeTagger wird darüber hinaus auch für die Lemmatisierung der russischen Sub-Korpora verwendet. Eine Schwierigkeit besteht dabei darin, dass das Vollformenlexikon, dass der *tagger* für die Lemmatisierung verwendet, notwendigerweise nicht alle in den russischen Sub-Korpora vorkommenden lexikalischen Einheiten enthält – ein Problem, das sich insbesondere negativ auf die Lemmatisierung von Fachtermini auswirken kann. Um die Lemmatisierung russischer Daten zu verbessern, wird in vorliegender Arbeit zusätzlich zu TreeTagger ein wiederum von Sharoff entwickelter *wrapper*[78] (vgl. Sharoff/Nivre 2011) für die Nutzung des Lemmatisierungswerkzeugs CSTlemma[79] (Jongejan/Dalianis 2009) auf russischen Daten eingesetzt. Dieses Werkzeug lernt Lemmatisierungsregeln und verwendet diese für die Lemmatisierung auch solcher Einheiten, die nicht in einem Lexikon hinterlegt sind.

Für das Wortarten-*tagging*, die morphologische Analyse und die Lemmatisierung der deutschen Sub-Korpora wird hingegen mit den mate-Werkzeugen[80] ein Teil einer vollständigen Annotations-*pipeline* verwendet (Björkelund et al. 2010). Die Wortartenannotation und die Lemmatisierung beruhen hier auf einem Verfahren des maschinellen Lernens; für die Lemmatisierung deutscher Wörter berichten die Autoren von einer Genauigkeit (*accuracy*) von 98,28 (ebd.: 34). Für das Training der deutschen Modelle wurde das Tiger-Korpus[81] (Brants et al. 2002) verwendet.

Abbildung 10 stellt die gesamte von mate angebotene Annotations-*pipeline* dar (Björkelund et al. 2010: 34). Anzumerken ist dabei, dass die Werkzeuge für die Annotation semantischer Rollen bislang nur für das Englische und Chinesische, nicht aber das Deutsche, zur Verfügung stehen.

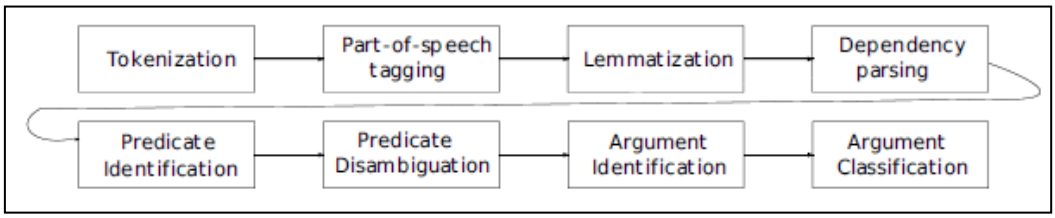


Abbildung 10: mate-Annotations-Pipeline nach Björkelund et al. (2010: 34)

4.4.2.2 Syntaktische Analyse

Die Analyse syntaktischer Abhängigkeiten in den Sub-Korpora erscheint für die Beschreibung und Extraktion wissenshaltiger Kontexte aus mehreren Gründen notwendig. So stellt einerseits die Analyse syntaktischer Dependenzen (Mel'čuk 1988) die syntaktischen Beziehungen zwischen Wörtern in Form von Dominanzbeziehungen dar und kann daher als „[...] especially suited to languages with free word order" wie etwa das Russische und tendenziell auch das Deutsche angesehen werden, „[...] since it directly represents the function of a word with respect to its head,

[76] Multilingual Text Tools and Corpora for Eastern and Central European Languages, http://nl.ijs.si/ME/, Zugriff am 24.03.2014.

[77] Die Spezifikation für das russische *tagset* findet sich hier: http://corpus.leeds.ac.uk/mocky/msd-ru.html, Zugriff am 24.03.2014.

[78] http://corpus.leeds.ac.uk/mocky/lemma-ru.tgz, Zugriff am 24.03.2014.

[79] http://cst.dk/download/uk/, Zugriff am 24.03.2014.

[80] http://code.google.com/p/mate-tools/, Zugriff am 24.03.2014.

[81] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html, Zugriff am 24.03.2014.

and it can capture discontinous structures like extraposition [...]“ (Seeker/Kuhn 2012: 3132). Andererseits kann gerade bei der Untersuchung wissenshaltiger Kontexte in *web*-Korpora nicht von einer Kontiguität im Sinne des direkten Nebeneinanderliegens der einzelnen Kontextbestandteile ausgegangen werden, vielmehr ist anzunehmen, dass KRC-Elemente durch für die Beschreibung und Suche nur wenig relevante Zeichenketten wie etwa Anführungszeichen, Einschübe und Verweise unterbrochen werden. Suchanfragen sind dann nicht auf der Ebene kontingenter *strings*, sondern auf der Ebene syntaktischer Einheiten und der zwischen ihnen bestehenden Beziehungen zu formulieren, wofür die syntaktische Dependenzanalyse eine wesentliche Voraussetzung ist, da andernfalls nicht sicherzustellen ist, dass die für die Musterbildung verwendeten Elemente tatsächlich in einer linguistischen Beziehung zueinander stehen. Entsprechend kann Walter (2010: 187) zeigen, dass die Verwendung dependenzbasierter Suchmuster bei der Extraktion von Definitionen aus deutschen Urteilstexten in deutlich besseren Präzisionswerten bei allerdings schlechterer Abdeckung resultiert als die Verwendung sequenzbasierter Muster, insgesamt jedoch in einem verbesserten f-Maß.

Aus praktischer Sicht sind bei der Auswahl geeigneter *parser* für Annotation russischer und deutscher Web-Korpora noch einige zusätzliche Anforderungen zu beachten. Hier ist in erster Linie die Forderung einer hohen Robustheit zu nennen, die es erlaubt, auch komplexe Sätze mit vertretbarem Aufwand zu analysieren. Dabei sind einerseits typische Eigenschaften fachsprachlicher Texte wie etwa eine hohe Komplexität der Satzstruktur, die Verwendung von Terminologie, spezifischen Abkürzungen sowie Hervorhebungen und Textstrukturierungen durch Nummerierung etc. (für das Deutsche vgl. Walter 2010: 128) und andererseits bereits benannte internettypische Aspekte wie beispielsweise sprachliche Fehler oder aber Mängel im Ergebnis der Dateikonvertierung (fehlerhafte Silbentrennung, Rechtschreib- oder typographische Fehler) ausschlaggebend für die Komplexität des Analyseprozesses. Eine weitere Anforderung war die Option, die Ausgabe des Parsers im CoNLL-Ausgabeformat (Buchholz/Marsi 2006: 151) zu erhalten, um für beide Sprachen eine einheitliche Struktur und damit einheitliche Verarbeitungsmethoden zu gewährleisten. Der Vorteil der CoNLL-Repräsentation besteht dabei darin, linguistische Annotationen auf unterschiedlichen Ebenen explizit in einem strukturierten Datenformat darzustellen. Darüber hinaus wurde solchen *parsern* der Vorzug gegeben, die in eine Annotations-*pipeline* integrierbar sind bzw. bereits in eine solche integriert wurden, da dies eine einheitliche Annotation der Daten in einer Gesamtstruktur mit aufeinander optimierten Verarbeitungsschritten ermöglicht und entsprechend eine Minimierung von Verarbeitungsfehlern verspricht.

4.4.2.2.1 Deutsch

Für die syntaktische Analyse deutscher Sätze wird der graph-basierte *parser* aus der mate-Toolsuite (Bohnet 2010) verwendet. Bohnet (ebd.: 96) berichtet von einem *labeled attachment score*[82] von 88,06, der bei der syntaktischen Analyse deutscher Sätze mit diesem *parser* erreicht wurde, Seeker und Kuhn (2012: 3138) erreichen auf ihren Testdaten einen *labeled attachment score* von 90,30. Bei dem verwendeten *parser* handelt es sich um ein statistisches Werkzeug, das ebenfalls auf dem Tiger-

[82] Dieser Wert gibt den Prozentsatz derjenigen *token* an, für welche der syntaktische Kopf und das Dependenz-Label korrekt annotiert wurden (Buchholz/Marsi 2006: 150).

Korpus (Brants et al. 2002) trainiert wurde[83]. Das Korpus umfasst insgesamt 35.000 Sätze aus deutschen Zeitungsartikeln (ebd.). Das Tiger-Annotationsschema baut auf dem des Negra-Korpus (Skut et al. 1997) auf und kodiert entsprechend Informationen auf drei Ebenen, und zwar auf den Ebenen des Einzelworts (Wortart), der syntaktischen Kategorien (z. B. Satz, Verbalphrase etc.) und der grammatischen Kategorien (grammatische Funktion wie z. B. Subjekt, vgl. Kübler 2010: 498). Abbildung 11 zeigt einen Tiger-Beispielsatz aus dem Bericht von Zinsmeister et al. (2001: 9) mit der Annotation auf mehreren Ebenen.

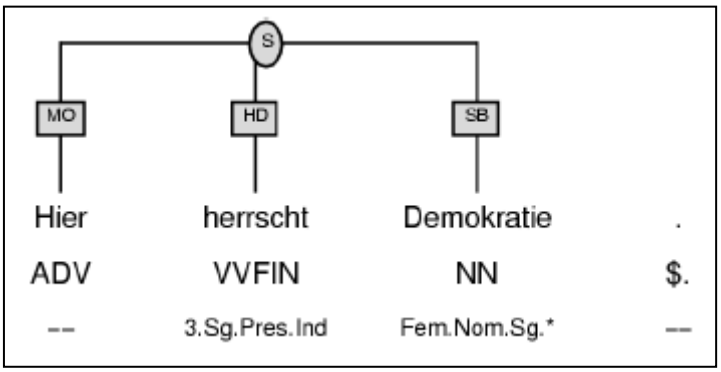


Abbildung 11: Tiger-Beispielsatz (Zinsmeister et al. 2001: 9)

Tabelle 35 stellt einen mit Hilfe der mate-Werkzeuge annotierten Beispielsatz aus dem Sub-EKD in dem von dem *parser* ausgegebenen CoNLL-Format in vereinfachter Form dar. Leere *tags* sind in der Darstellung nicht berücksichtigt. Spalte 1 enthält dabei die Identifikationsnummer (ID) für jedes *token* (Wort) des Satzes, die Spalten 2-5 die in den vorhergehenden Annotationsstufen gewonnenen Informationen zu Wortform, Lemma, Wortart sowie Morphologie. Schließlich verweist Spalte 6 auf die ID desjenigen *tokens*, das als syntaktischer Kopf des betreffenden Wortes fungiert (syntaktische Dependenz), und Spalte 7 benennt die grammatische Funktion des *tokens* und entspricht damit den Kantenlabeln aus dem Tiger-Beispielsatz in Abbildung 11. „NK“ bedeutet dabei *noun kernel* und bezeichet ein Element einer Nominalphrase. „SB“ bezeichnet das Subjekt des Satzes. „PAR“ bezeichnet eine Parenthese, wohingegen „MO“ eine Modifikationsrelation zwischen *token* und syntaktischem Kopf beschreibt. „PD“ bezeichnet schließlich den Prädikativ.

ID	Wortform	Lemma	POS	Morphologie	Dep	Funk
1	Eine	ein	ART	case=nom|number=sg|gender=fem	2	NK
2	Windenergieanlage	windenergieanlage	NN	case=nom|number=sg|gender=fem	6	SB
3	(	--	$(	_	2	--
4	Windkraftanlage	windkraftanlage	NN	case=nom|number=sg|gender=fem	2	PAR
5	)	--	$(	_	4	--
6	ist	sein	VAFIN	number=sg|person=3|tense=pres|mood=ind	0	--
7	nur	nur	ADV	_	8	MO

[83] Das Tiger-Annotationsschema für die syntaktische Annotation ist hier einsehbar: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-syntax.pdf, Zugriff am 24.03.2014.

8	scheinbar	scheinbar	ADJD	degree=pos	6	MO
9	eine	ein	ART	case=nom|number=sg| gender=fem	11	NK
10	einfache	einfach	ADJA	case=nom|number=sg| gender=fem|degree=pos	11	NK
11	Konstruktion	konstruktion	NN	case=nom|number=sg| gender=fem	6	PD
12	.	--	$.	_	11	--

Tabelle 35: Beispielsatz aus Sub-EKD nach Analyse mit den mate-Werkzeugen

4.4.2.2.2 Russisch

Für die syntaktische Analyse russischer Sätze stehen weniger Ressourcen zur Verfügung als für das Deutsche. Allerdings existiert mit SynTagRus[84] (Nivre et al. 2008) auch für das Russische ein syntaktisch annotiertes Korpus, auf dessen Grundlage statistische Analysewerkzeuge trainiert werden können. Experimente zur syntaktischen Analyse russischer Texte wurden mit dem statistischen Parser MaltParser[85] (Nivre et al. 2007) durchgeführt (Sharoff/Nivre 2011). Darüber hinaus kann eine von Serge Sharoff entwickelte *pipeline*[86] verwendet werden, die den TreeTagger und den *wrapper* für CSTlemma integriert und das erwähnte, bereits trainierte Parsingmodell[87] nutzt.

SynTagRus enthält literarische sowie populärwissenschaftliche und politisch-publizistische Texte aus den Jahren 1980-2008 sowie Nachrichtentexte (Pamjatka 2011). Zum Zeitpunkt der ersten Experimente mit Maltparser im Jahr 2008 enthielt das Korpus etwa 460.000 Wörter (Nivre et al. 2008: 642). Die linguistische Annotation des Korpus‘ wurde automatisiert mit Hilfe des Etap-Parsers (Iomdin et al. 2012) durchgeführt und manuell korrigiert sowie vollständig disambiguiert (Pamjatka 2011). Die der Annotation zugrundeliegende linguistische Theorie ist Mel’čuks *Meaning Text Theory* (Теория Смысл ⟷ Текст, vgl. Mel’čuk et al. 1995).

Die im Korpus enthaltenen Wörter sind feinkörnig morphologisch annotiert. Die Annotation markiert syntaktische Dependenzen und beschreibt dabei rund 65 syntaktische Beziehungen (Pamjatka 2011). Einige der wichtigsten syntaktischen Relationen werden in Tabelle 36 beschrieben (vgl. ebd., vgl. Nivre et al. 2008: 642 f.). Spalte 5 gibt dabei zusätzlich zu der im Beispiel dargestellten Realisierungsform der fraglichen Beziehung weitere Umsetzungsmöglichkeiten an.

Russische Bezeichnung	**Tag**	**Beschreibung**	**Beispiel**	**Weitere Realisierungsformen**
Предикативное синтаксическое отношение	предик	markiert die Beziehung zwischen Subjekt [Y] und Prädikat [X]	Петя [Y] читает [X] *(Petja [Y] liest [X])*	Passivierung, verblose Konstruktionen mit oder ohne Bindestrich, Partizipialkonstruktionen, unpersönliche Konstruktionen, Infinitivsätze, Hypotaxe

[84] http://www.ruscorpora.ru/en/, Zugriff am 24.03.2014.
[85] http://www.maltparser.org/, Zugriff am 24.03.2014.
[86] http://corpus.leeds.ac.uk/mocky/russian-malt.tgz/, Zugriff am 24.03.2014.
[87] http://corpus.leeds.ac.uk/mocky/rus-test.mco/, Zugriff am 24.03.2014.

Агентивное синтаксическое отношение	агент	markiert die Beziehung zwischen einem Verb bzw. einer Nominalisierung [X] und einer das Agens der Handlung wiedergebenden Nominalgruppe im Instrumental [Y]	прием [X] президентом [Y] представителей Думы *(der Empfang [X] der Duma-Repräsentanten durch den Präsidenten [Y])*	-
Квазиагентивное синтаксическое отношение	квазиагент	markiert die erste Valenzposition [Y] des logischen Prädikats (*predikativnoe slovo*) [X]	приход [X] друга [Y] *(die Ankunft [X] des Freundes [Y])*	Präpositionalphrasen
Комплетивные синтаксические отношения	{1-компл, 2-компл, ..., 5-компл}	markiert weitere Valenzpositionen [Y] des logischen Prädikats (*predikativnoe slovo*) [X]	Я вас [Y1] предупреждал [X], чтобы [Y2] вы не торопились. *(Ich habe euch [Y1] gesagt [X], dass [Y2] ihr euch nicht beeilen sollt.)*	-
Присвязочное синтаксическое отношение	присвяз	markiert Kopulaverben [X], die das Subjekt mit dem Prädikativ [Y] verbinden	Петя казался [X] больным [Y]. (*Petja schien [X] krank zu sein [Y].)*	-
Орпределительное синтаксическое отношение	опред	markiert die Beziehung zwischen einem nominalen oder adjektivischen Kopf [X] und einem abhängigen Adjektiv oder Partizip [Y]	вижу красивый [Y] дом [X] *(ich sehe ein schönes [Y] Haus [X])*	-
Предложное синтаксическое отношение	предл	markiert die Beziehung zwischen einer Präposition [X] und einer Nominalgruppe [Y]	между [X] Сциллой [Y] и Каридбой *(zwischen [X] Scylla und Charybdis [Y])*	-

Tabelle 36: Zentrale syntaktische Relationen in SynTagRus

Tabelle 36 gibt einen Überblick über die Annotation grundlegender syntaktischer Beziehungen in SynTagRus. Im Vergleich zu dem bereits diskutierten Tiger-Annotationsschema fällt insbesondere der

Verzicht auf eine stringente Unterscheidung zwischen syntaktischen Kategorien und grammatischen Funktionen zugunsten einer konsistenten Modellierung syntaktischer Beziehungen im Sinne der *Meaning Text Theory* ins Auge. Tabelle 37 stellt die Ausgabe des russischen *parsers* nach Durchlaufen der gesamten Annotationspipeline anhand eines Beispielsatzes dar. Auch hier wurden leere oder redundante *tags* in der Darstellung nicht berücksichtigt. ROOT bezeichnet die Satzwurzel, PUNC markiert Interpunktionszeichen. Zu beachten ist die fehlerhafte Lemmatisierung der beiden Adjektive.

ID	Wortform	Lemma	POS	Morphologie	Dep	Funk
1	Проверьте	проверить	V	Vmm-2p-a-e	0	ROOT
2	состояние	состояние	N	Ncnsnn	1	1-компл
3	резиновых	резиновых	A	Afp-pgf	5	опред
4	опорных	опорных	A	Afp-pgf	5	опред
5	прокладок	прокладка	N	Ncfpgn	2	квазиагент
6	пружин	пружина	N	Ncfpgn	5	квазиагент
7	.	.	S	SENT	6	PUNC

Tabelle 37: Beispielsatz aus Sub-EKR nach Analyse mit Hilfe der Annotationspipeline

4.4.3 Formatierung

Nach der linguistischen Annotation werden die Korpora schließlich in ein Datenformat überführt, das die einfache Weiterverarbeitung der Daten im Sinne einer linguistischen Analyse sowie von Extraktionsexperimenten erlaubt. Hierfür wird XML (*Extensible Markup Language*)[88] gewählt. Der Vorteil von XML beispielsweise gegenüber den von den *parsern* ausgegebenen tabulator-getrennten Formaten besteht insbesondere darin, dass es sich bei XML um ein strukturiertes Datenformat handelt, das hierarchisch aufgebaut ist und mithin den gezielten Zugriff auf Elemente innerhalb der XML-Baumstruktur erlaubt. Darüber hinaus stehen für die Arbeit mit XML über das CPAN-Portal[89] Perl-Module für eine Vielzahl unterschiedlichster Anforderungen bereit. Nicht zuletzt besteht bei der Nutzung von XML auch die Möglichkeit, die Datei auf ihre Übereinstimmung mit einer sogenannten Dokumenttypdefinition (DTD) hin zu überprüfen (zu validieren) und auf diese Weise auf etwaige Inkonsistenzen zu überprüfen. Eine DTD legt die hierarchische Struktur eines XML-Dokuments fest und definiert die zu einem Element des XML-Baums gehörigen Attribute. Abbildung 12 stellt die einfache DTD dar, die für die Formatierung der Sub-Korpora verwendet wird:

[88] Auch für die Spezifikation von XML existiert eine Empfehlung des *World Wide Web Consortium*: http://www.w3.org/TR/2008/REC-xml-20081126/, Zugriff am 24.03.2014.

[89] http://www.cpan.org/, Zugriff am 24.03.2014.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- anno.dtd -->
<!ELEMENT corpus (sentence+)>
<!ELEMENT sentence (token+)>
<!ELEMENT token (#PCDATA)>
<!ATTLIST corpus name CDATA #REQUIRED>
<!ATTLIST sentence sent_id CDATA #REQUIRED>
<!ATTLIST token
        id CDATA #REQUIRED>
        lemma CDATA #REQUIRED
        isterm (true|false) #REQUIRED
        pos CDATA #REQUIRED
        morph CDATA #REQUIRED
        dep CDATA #REQUIRED
        sfun CDATA #REQUIRED
>
```

Abbildung 12: DTD für die Sub-Korpora

Der Abbildung ist zu entnehmen, dass die Sub-Korpora auf drei Ebenen beschrieben werden, und zwar der Ebene des Korpus‘, der Ebene einzelner Sätze sowie der Ebene einzelner Wörter (*tokens*). Hierbei handelt es sich um Elemente des durch die DTD definierten XML-Baums, die in einer hierarchischen Beziehung zueinander stehen, d. h. das Korpus besteht aus mindestens einem Satz (<!ELEMENT corpus (sentence+)>), der wiederum aus mindestens einem *token* besteht (<!ELEMENT sentence (token+)>). Dem Korpus ist das obligatorische Attribut „name“ zugeordnet (<!ATTLIST corpus name CDATA #REQUIRED>), das angibt, welches Sub-Korpus in einer XML-Datei enkodiert wurde. Zudem ist jeder Satz im Korpus anhand des obligatorischen Attributs „sent_id“, einer individuellen Identifikationsnummer, eindeutig identifizierbar (<!ATTLIST sentence sent_id CDATA #REQUIRED>). Für jedes *token* übernimmt die XML-Datei aus der *parser*-Ausgabe die folgenden Informationen:

- eine Identifikationsnummer, welche der Position des *tokens* im Satz entspricht (<!ATTLIST token id CDATA #REQUIRED>),
- das Lemma (<!ATTLIST token lemma CDATA #REQUIRED>),
- den Wortarten-*tag* (<!ATTLIST token pos CDATA #REQUIRED>),
- morphologische Informationen (<!ATTLIST token morph CDATA #REQUIRED>),
- den syntaktischen Kopf (<!ATTLIST token dep CDATA #REQUIRED>),
- die syntaktische Funktion (<!ATTLIST token sfun CDATA #REQUIRED>).

Zusätzlich zu diesen Angaben wird bei der Konvertierung der geparsten Dateien nach XML eine Termerkennung durchgeführt. Diese beruht auf einem einfachen *string*-Abgleich der eingegebenen Lemmata mit den ebenfalls lemmatisierten Zieltermini des Goldstandards. Die Zuordnung von Zieltermini zu den KRCs des Goldstandards wurde manuell auf Grundlage der den Annotatoren bereits vorgelegten Zielterminus-Kandidaten durchgeführt, wobei im Goldstandard vorkommende Varianten ebenfalls berücksichtigt wurden, um eine möglichst gute Abdeckung zu erreichen (s. o.). Entspricht ein *token* einem der im Goldstandard annotierten Zieltermini, wird dies durch das obligatorische *token*-Attribut „isterm=‘true‘“ ausgedrückt. Im Ergebnis werden in jedem Satz diejenigen *strings* als Termini gekennzeichnet, die in mindestens einem KRC des Goldstandards als

Zielterminus fungieren, auch wenn sie in dem betreffenden Satz selbst nicht Zielterminus sind. Abbildung 13 stellt die *token*-Annotation für einen Zielterminus am Beispiel des zweistelligen Terminus „taktische U-Boote“ aus dem Sub-TKD dar:

```
<token id="6" lemma="taktisch" isterm="true" pos="ADJA"
morph="case=nom|number=pl|gender=fem|degree=pos" dep="7"
sfun="NK">taktische</token><token id="7" lemma="u-boot" isterm="true" pos="NN"
morph="case=nom|number=pl|gender=fem" dep="3" sfun="PAR">U-Boote</token>
```

Abbildung 13: Token-Annotation im XML-Format

Mit diesem Format steht sowohl für die deutschen als auch für die russischen Korpora eine einheitliche Darstellung zur Verfügung, so dass für die weitere Analyse und Verarbeitung der Daten auf sprachspezifische Anpassungen auf Programmebene verzichtet werden kann. Da es sich indes bei der Term-Erkennung um eine selbstentwickelte Annotationskomponente handelt, soll diese kurz evaluiert werden. Dafür werden für jedes Sub-Korpus jeweils 100 Sätze des Goldstandards auf die Präzision und Vollständigkeit der Term-Erkennung überprüft. Unvollständig annotierte Termini gehen zwar als korrekt annotierte Termini in die Bewertung ein, da zumeist ein attributiv gebrauchtes Adjektiv nicht als Teil des Terminus‘ erkannt, aber zumindest dessen syntaktischer Kopf korrekt annotiert wurde. Unvollständige Terminus-Annotationen werden aber im Sinne einer transparenten Darstellung des Ergebnisses gesondert ausgewiesen. Tabelle 38 fasst die Ergebnisse zusammen.

	Sub-EKD	**Sub-TKD**	**Sub-EKR**	**Sub-TKR**
Erkannte Termini	200 Termini, davon 12 Zweiworttermini und 2 Dreiworttermini	145 Termini, davon 11 Zweiworttermini, 4 Dreiworttermini und 2 Vierworttermini	232 Termini, 94 Zweiworttermini, 31 Dreiworttermini und 10 Vierworttermini	150 Termini, davon 45 Zweiworttermini, 10 Dreiwort-termini, 3 Vierworttermini und 2 Fünfworttermini
Unvollständig erkannte Termini	-	-	5 Termini	5 Termini
Präzision	1,00	0,99	1,00	1,00
Abdeckung	0,97	0,95	0,90	0,79

Tabelle 38: Evaluation der Term-Erkennung

Tabelle 38 zeigt wesentliche Unterschiede in der terminologischen Charakteristik der deutschen und russischen Sub-Korpora auf. Zunächst weisen die russischen Korpora im Vergleich zu ihren deutschen Pendants jeweils einen deutlich höheren Anteil an Mehrworttermini auf, ein Umstand, der darauf zurückzuführen ist, dass im Russischen im Vergleich zum Deutschen das Wortbildungsverfahren der Komposition wesentlich seltener als Möglichkeit der Bildung komplexer Benennungen Verwendung findet. Des Weiteren kann festgestellt werden, dass das Gros der Fehler bei der Term-Erkennung entweder auf einer inkonsistenten Lemmatisierung, d. h. Unterschieden in der Lemmatisierung

zwischen der Liste mit den Zieltermini und den Termini in den Korpora, oder aber auf echten Lemmatisierungsfehlern, d. h. einer generell inkorrekten Lemmatisierung beispielsweise einer Genitiv- oder Pluralform, beruht. Insbesondere für das Russische mit seiner stark ausgeprägten Flexionsmorphologie scheint die Lemmatisierung der Wörter trotz der Nutzung von CSTlemma noch problematisch, wodurch sich zumindest ein Teil der deutlich niedrigeren Abdeckung bei der Term-Erkennung auf den russischen Korpora erklären lässt.

4.5 Zusammenfassung

In diesem Kapitel wurde beschrieben, auf welche Weise und unter Zuhilfenahme welcher Werkzeuge die Datengrundlage erstellt wurde, die in den folgenden Kapiteln analysiert werden soll. In der Einleitung waren für dieses Kapitel zwei Zielstellungen formuliert worden, und zwar:

- dass die Extraktionskorpora unterschiedliche Fachgebiete abdecken sollen,
- dass die Annotationsstudie nicht nur die Grundlage für die in den nächsten Kapiteln durchzuführenden Analysen bilden, sondern auch Aufschluss darüber gehen soll, in welchem Maße Übersetzer überhaupt einen Konsens in Bezug auf die Nützlichkeit, d. h. den KRC-Status von Kandidaten abbilden können.

Im weiteren Verlauf dieses Kapitels waren im Hinblick auf die Dokumentation und Charakterisierung der erstellten Korpora weitere Fragen aufgeworfen worden:

- Sind die automatisch gecrawlten Korpora fachlich spezialisiert?
- In welchen Textsorten treten wissenshaltige Kontexte überwiegend auf?

Mit Hilfe des Crawlers „Babouk“ waren zunächst für das Russische und das Deutsche jeweils zwei *web*-Korpora erstellt worden, die auf iterativ und manuell erstellten Terminuslisten beruhen. Die durchgeführten Analysen haben gezeigt, dass zwischen diesen Korpora durchaus inhaltliche Überschneidungen existieren, die allerdings bei der Erstellung der für die weiteren Analysen zu verwendenden Sub-Korpora weitgehend ausgeräumt werden konnten (Tabelle 33). Im Hinblick auf die Fragestellung, ob es sich bei den automatisch gesammelten Texten tatsächlich auch um Fachtexte handele, wurde zunächst festgestellt, dass „Babouk“ Texte nach der Häufigkeit des Auftretens bestimmter Termini selektiert. In diesem Sinne sind folglich alle automatisch gecrawlten Texte – also auch Listen, Forumsbeiträge etc. – tatsächlich Fachtexte. Da eine solche Charakterisierung für das hier verfolgte Vorhaben jedoch nicht ausreichend ist, wurde auf Makroebene eine Charakterisierung der Großkorpora durchgeführt. Zu diesem Zweck wurden einige der von Teich und Fankhauser (2010) genannten korpuslinguistischen Fachlichkeitskriterien – *type-token*-Verhältnis sowie relative Frequenzen von Adverbien und Nomina – ebenso eingesetzt wie Untersuchungen der durchschnittlichen Satzlänge. Im Ergebnis zeigten sich für die TTR nur schwache Unterschiede bzw. für eines der russischen Korpora konnte die von Teich und Fankhauser formulierte Hypothese gar nicht bestätigt werden, womit – in Anbetracht dessen, dass das als Referenz verwendete RIC einen mittleren Fachlichkeitsgrad aufweist, die TTR aber in der Literatur kontrovers diskutiert wird – keine eindeutigen Schlussfolgerungen in Bezug auf den Fachlichkeitsgrad der gecrawlten Korpora gezogen werden können. Andererseits konnten ein stärker ausgeprägter Nominalstil sowie geringere Anteile von Adverbien, jedoch wiederum nur relativ kurze Satzlängen festgestellt werden. Im Großen und Ganzen scheint diese Makroanalyse darauf hinzudeuten, dass die automatisch erstellten Großkorpora eine Mischung unterschiedlichster Texte darstellen, die nicht alle im klassischen Sinne als „Fachtexte“ bezeichnet werden können, wobei allerdings die Bedeutung dieses Begriffs für *web*-Textsorten selbst nicht klar ist. Bestenfalls scheint in

Bezug auf die Frage der Fachlichkeit der „Babouk“-Korpora eine Abschätzung nach unten möglich zu sein, und zwar dahingehend, dass sich Letztere durch typische Merkmale wie etwa einen ausgeprägteren Nominalstil von nicht-fachlichen Korpora unterscheiden und mithin gewisse Spezialisierungsmerkmale aufweisen.

Die Auswahl von Texten für die Sub-Korpora des Goldstandards scheint diesbezüglich einige Mängel der Großkorpora ausgeglichen zu haben – werden etwa die durchschnittlichen Satzlängen von Groß- und Subkorpora verglichen (Tabellen 9 und 32), so fällt auf, dass erstere in den Subkorpora leicht erhöht sind. Auch wurden überwiegend solche Texte für die Übernahme in den Goldstandard ausgewählt, welche gemäß der entwickelten *ad-hoc*-Textsortenklassifikation als „erklärend“ charakterisiert wurden, womit in der Literatur genannte Hypothesen zu textsortentypischen Verteilungen wissenshaltiger Kontexte bestätigt zu werden scheinen.

Hinsichtlich der Frage, ob Übersetzer in Bezug auf wissenshaltige Kontexte überhaupt einen Konsens erreichen können, wurde – mit Vorbehalten – festgestellt, dass zumindest einige linguistische Indizien darauf hinzudeuten scheinen, die mit der Bewertung von KRC-Kandidaten beauftragten Übersetzer jedoch insbesondere bei Zweifels- und Grenzfällen unterschiedliche Annotationsstrategien anwenden. Insgesamt wurde eine recht umfangreiche Spezifikation wissenshaltiger Kontexte entwickelt, darüber hinaus wurden Desiderata für zukünftige Studien formuliert. Nicht zuletzt wurden Vorverarbeitungs- und Annotationsschritte für die Subkorpora beschrieben.

5. Linguistische Eigenschaften wissenshaltiger Kontexte

5.1 Einführung

In diesem Kapitel soll anhand der in den Goldstandard übernommenen russischen und deutschen KRCs ein Überblick über linguistische Eigenschaften wissenshaltiger Kontexte erarbeitet werden. Für eine grobe Gliederung der Untersuchung bzw. die Identifikation relevanter Merkmalskategorien soll dabei zunächst auf die Ergebnisse bereits existierender Studien zum Deutschen und zu anderen Sprachen zurückgegriffen werden, deren Ergebnisse mittels der hier unternommenen empirischen Untersuchung überprüft und ergänzt werden sollen. Die aus der Literatur entnommenen Untersuchungsmerkmale sollen alsdann durch solche Merkmale ergänzt werden, die sich an der systemisch-funktionalen Linguistik (SFL, Halliday 1994[2]) und Ansätzen zur Beschreibung linguistischer Verfahren der Verallgemeinerung (Drescher 1992) orientieren.

Eine solche Untersuchung linguistischer Merkmale wissenshaltiger Kontexte erscheint dabei in zweierlei Hinsicht als relevant. Einerseits kann sie – als Kombination quantitativer und qualitativer Vorgehensweisen – Hinweise auf stilistische Besonderheiten und damit Antworten auf die Frage geben, ob wissenshaltige Kontexte im Vergleich zu anderen Aussagetypen durch besondere stilistische Eigenheiten gekennzeichnet sind und ob möglicherweise ein eigener Funktionalstil des Erklärens und Definierens postuliert werden kann. Im Vergleich zum bisherigen Wissensstand zu wissenshaltigen Kontexten und Definitionen stellt die empirisch fundierte Diskussion dieser Frage einen wesentlichen Fortschritt dar, da bisherige Arbeiten ein bruchstückhaftes, unvollständiges Bild wissenshaltiger Kontexte zeichnen. So merkt etwa Walter (2010: 40), wie bereits dargestellt, das Fehlen systematischer statistischer Untersuchungen an und findet in seinen eigenen Daten ein weitaus größeres Variationsspektrum vor, als es in der Literatur beschrieben wird. Für vorliegende Untersuchung, die sich mit der im Vergleich zu Walters Untersuchungsgegenstand – Definitionen juristischer Begriffe in Urteilsbegründungen – deutlich weiter gefassten Kategorie des wissenshaltigen Kontexts befasst, kann daher mit einem noch breiteren Variationsspektrum gerechnet werden. Empirisch fundierte Beschreibungen wissenshaltiger Kontexte für das Deutsche und Russische fehlen dabei völlig.

Die Untersuchung linguistischer Merkmale wissenshaltiger Kontexte ist aber auch insofern von Interesse, als – sollen moderne computerlinguistische Verfahren, beispielsweise Methoden des maschinellen Lernens für die automatische Identifikation erklärender und definierender Aussagen in Textkorpora eingesetzt werden – die sprachlichen Merkmale von KRCs im Vorfeld möglichst genau zu beschreiben sind, so dass für die automatische Identifikation von KRCs ausreichend diskriminatorische Merkmale zur Verfügung stehen. Die Kombination qualitativer mit quantitativen Methoden erscheint dabei insofern als sinnvoll, als Letztere den Korpuscharakter des erstellten Goldstandards durch Anwendung automatischer Annotationen und statistischer Analysen sinnvoll ausnutzen können. Quantitative Daten können so mit Referenzdaten verglichen und hinsichtlich ihrer statistischen Signifikanz (für eine Erläuterung dieses Begriffs vgl. Sachs/Hedderich 2009[13]: 361 ff.) und der Stärke des ermittelten Effekts (vgl. hierzu Rosenthal et al. 2000: 4) bewertet werden und tragen somit wesentlich zu einer systematischen Charakterisierung wissenshaltiger Kontexte bei. Im quantitativen Sinne wird Stil dabei als „probabilistisches Konzept“ (vgl. Tuldava 2005: 369) aufgefasst und entsteht als Resultat der Wahl eines kompetenten Sprechers aus spezifischen

funktionalstilistischen Sprachmitteln. Mehler (2005: 339, sic!) fasst diesen Gedanken prägnant zusammen:

> „Die Kernthese der quantitativen Stilistik lautet, daß stilistische Unterschiede von Autoren, Textgruppen oder Textsorten unter Rekurs auf die quantitativen Eigenschaften der betroffenen Texte gemessen werden können […]. Sie geht von der Annahme aus, daß Stilunterschiede anhand von (nicht-deterministischen, nicht-regelbasierten) Tendenzen zur Verwendung von Ausdrucksarten beobachtet und mit Hilfe von Wahrscheinlichkeitsverteilungen beschrieben werden können. In diesem Sinne betrachtet die quantitative Stilistik den probabilistischen Charakter textueller Zusammenhänge als primäre Eigenschaft stilistischer Phänomene […]. Die Operationalisierung des quantitativen Stilbegriffs setzt die Spezifikation jener Texteigenschaften voraus, die zur Differenzierung des Stils einzelner Autoren oder Autorengruppen dienen. […] Dabei werden nur solche Texteigenschaften, die in Abhängigkeit von pragmatischen Faktoren der Textproduktion variieren und folglich die stilistische, autorenspezifische Differenzierung von Texten erlauben, als quantitative Stilcharakteristiken eingestuft [...]. Der Stil eines Texts wird dann als die Gesamtheit seiner objektiv faßbaren, quantitativen Stilcharakteristiken definiert […].“

In diesem Sinne kombiniert die in diesem Abschnitt unternommene Untersuchung, wie bereits angemerkt, klassische qualitative Verfahren mit quantitativen Methoden der linguistischen Beschreibung und erhebt auch Merkmale wie Satzlänge oder syntaktische Komplexität, die für die qualitative Untersuchung von Stilmerkmalen nicht immer von primärem Interesse sind, allerdings im Vergleich zu Referenzdaten möglicherweise Hinweise für die Erstellung eines stilistisch-linguistischen Profils geben können. Qualitative Vorgehensweisen tragen dagegen dem begrenzten Umfang des Goldstandards Rechnung und beschreiben Phänomene, die automatisch nur schwer zu annotieren sind oder aber nur selten in den untersuchten Datensätzen auftreten. Im nächsten Abschnitt soll zunächst ein Katalog interessierender linguistischer Merkmale zusammengestellt werden.

5.2 Linguistische Fragestellungen

5.2.1 Ein Merkmalskatalog

5.2.1.1 Aus der relevanten Literatur entnommene Merkmale und Hypothesen

Die der zitierten Literatur entnommenen Hypothesen zu linguistischen Eigenschaften wissenshaltiger Kontexte sollen in diesem und den folgenden Abschnitten mittels des Einbezugs der SFL, aber auch anderer linguistischer Arbeiten systematisch ergänzt, operationalisiert und an allgemeinere, d. h. von der Ebene der konkreten linguistischen und statistischen Merkmale abstrahierende und einen größeren Erklärungszusammenhang herstellende Fragestellungen gekoppelt werden. Da der Literatur für das Russische keine eigenen Hypothesen entnommen werden konnten, wird dabei auf Grundlage der Annahme, dass KRCs einen eigenen Aussagetypus bzw. eine eigene Textsorte (vgl. dazu die Diskussion in Kapitel 3) konstituieren, von der Übertragbarkeit zumindest derjenigen linguistischen Hypothesen, die sich nicht auf stark spezialisierte Typen von Definitionen (wie im Falle Walters) bzw. Kontexten beziehen, auch auf das Russische ausgegangen. Ob eine solche Übertragbarkeit gegeben ist und ob mithin auf Grundlage empirischer Daten ein eigener Aussagetyp KRC postuliert werden kann, ist Teil des mit der hier unternommenen Analyse verfolgten Erkenntnisinteresses.
Aus der in Abschnitt 3.2 referierten Literatur ergeben sich zunächst zwei Gruppen zu untersuchender Merkmale, nämlich einerseits lexikalische und andererseits strukturelle und grammatische Merkmale. Das lexikalische Merkmal Boostwörter ergibt sich dabei direkt aus der zitierten Literatur

und soll dazu dienen, die dort referierten Thesen zu erhärten, zu erweitern oder zu widerlegen. Frequenzen sollen zudem auch auf der Ebene der Wortarten untersucht werden, um die gewonnenen Informationen zu ergänzen und zu erweitern. Entsprechend kann die Hypothese formuliert werden, dass die KRCs des Goldstandards einen relativ hohen Anteil an (generischen) Nomina (nach Walther 2010, Cramer 2011) aufweisen müssten. Des Weiteren kann mit Cramer (2011) von einem relativ häufigen Gebrauch lexikalischer Generalisierungssignale und bestimmter definitionstypischer Adverbien ausgegangen werden. Die Untersuchung von Stopwörtern ergibt sich komplementär dazu als negatives Merkmal. Nicht zuletzt sollten die in der Literatur angeführten *knowledge patterns* ebenfalls bereits auf dieser Analyseebene in Erscheinung treten.

Auf strukturell-grammatischer Ebene scheinen zunächst Pearsons (1998) Aussagen zum Artikelgebrauch die Formulierung spezifischer Hypothesen für das Deutsche (das Russische verfügt ohnehin nicht über Artikel) kaum zuzulassen[90], da der Artikelgebrauch im Deutschen deutlich von den englischen Normen abweicht. Das von Walter (2010) genannte Merkmal des Auftretens von Oberbegriffen des Definiendums wird im Rahmen der vorliegenden Arbeit nicht untersucht, da die Implementierung eines entsprechenden Werkzeugs insbesondere für das Russische über den Rahmen dieser Arbeit hinausgeht und eine manuelle Annotation als zu aufwändig erscheint. Dokumentbezogene Merkmale können zudem wegen der Eingrenzung wissenshaltiger Kontexte auf die Satzebene nicht ohne Weiteres berücksichtigt werden und stehen im Widerspruch zu der Darstellung Cramers (2011). Pearsons Aussage, dass sich die Explikation des Definiendums stets im Hauptsatz des KRCs finden muss, scheint indes mit Walters (2010) Feststellung, dass das Definiendum stets als Argument des Prädikats realisiert ist, kompatibel. Mit Pearson (1998) ebenso wie mit Cramer (2011) und Walter (2010) kann des Weiteren vom Auftreten von Geltungsrestriktionen und Fachgebietsangaben in Form spezifischer Phrasen ausgegangen werden. Nicht zuletzt erscheint die auch von Walter (ebd.) ermittelte Länge wissenshaltiger Kontexte als einfach zu erhebendes Merkmal.

Aus der einschlägigen Literatur ergeben sich entsprechend unten stehende Forschungsfragen, die in den folgenden Abschnitten weiter spezifiziert und ergänzt werden sollen:

- Lexikalische Merkmale
    - Welche Wortarten treten häufig bzw. selten auf?
    - Welche lexikalischen Einheiten (Boostwörter) treten besonders häufig auf, finden sich ebenfalls generische Nomina, definitionstypische Adverbien oder andere lexikalische Generalisierungssignale?
    - Welche lexikalischen Einheiten (Stopwörter) treten besonders selten auf?
    - Welche lexikalischen Einheiten kommen als *knowledge pattern* infrage?
- Strukturelle und grammatische Merkmale
    - In welcher grammatischen und syntaktischen Form ist das Definiendum realisiert?
    - Finden sich Geltungsrestriktionen auf phraseologischer Ebene?
    - Wie lang sind wissenshaltige Kontexte?

[90] Dies gilt auch insofern, als Pearson (1998: 158) für *semi-formal defining expositives* andere Regularitäten des Artikelgebrauchs beschreibt. Da Pearsons Unterscheidung zwischen *formal* und *semi-formal defining expositives* in vorliegender Arbeit nicht repliziert wird, scheint der Nutzen ihrer Beschreibung im Hinblick auf Artikel als fragwürdig. Generell scheint es sich bei den linguistischen Kriterien Pearsons eher um Heuristiken als um empirische Beschreibungen zu handeln.

5.2.1.2 Generalisierung

Zur Ergänzung und Erweiterung der bislang identifizierten Merkmalskategorien bzw. Forschungsfragen soll noch ein weiterer Ansatz genutzt werden, und zwar ein Ansatz zur Beschreibung generalisierender Aussagen. Die Generalisierbarkeit der durch den Kontext-Kandidaten vermittelten Information gehörte zu den Validitätskriterien bei der Annotation des Goldstandards, mithin besteht Grund zu der Annahme, dass sich neben den von Cramer (2011) genannten lexikalischen Generalisierungssignalen auch auf anderer Ebene Hinweise auf die Verallgemeinerbarkeit der Information finden.
Informationen darüber, mit welchen sprachlichen Mitteln Generalisierungen vorgenommen werden können, gibt etwa die Arbeit von Drescher (1992). Drescher untersucht am Beispiel des Französischen linguistische Verfahren der Generalisierung als Mittel der Textkonstitution und verweist zunächst unter Rückbezug auf frühere Arbeiten (ebd.: 53) darauf, dass neben lexikalischen Einheiten insbesondere auch morphologische und syntaktische Besonderheiten für generalisierbare Aussagen konstitutiv seien. Für das Thema Generalisierung von Interesse seien insbesondere die Realisierung der Verbgruppe sowie spezifische syntaktische Strukturen, wobei Parallelen zu häufigen für die Beschreibung von definitorischen Äußerungen genutzten Mustern nicht von der Hand zu weisen sind (ebd. 54):

> „Im Rahmen der grammatischen Generizitätsforschung werden bislang vor allem Standardsätze der Art *un, les, le* + Nominal (NP)- und Verbalphrase (VP) untersucht, wobei ausdrucksseitige Merkmale der Generizität insbesondere im Bereich der Nominalphrase in den Blick geraten. Besondere Beachtung finden die Kategorie der Determinanten und hier vor allem die Formen des Artikels […]. Wenig untersucht wurde bislang der generische Charakter bestimmter unpersönlicher Sätze, Passivkonstruktionen oder Präsentative […]. Zunehmend Beachtung finden generische Aspekte im Bereich der Verbalphrase. Damit geraten auch das Tempus des Verbs sowie einige Adverbien in den Blick.“

In Bezug auf den Gebrauch des Artikels, der beispielsweise von Pearson (1998) für das Englische thematisiert worden war, lassen Dreschers Aussagen zum generalisierenden Potential des Artikels die Formulierung klar voneinander abgegrenzter Hypothesen nicht zu: Sowohl der unbestimmte als auch der bestimmte Artikel scheinen in generalisierenden Aussagen möglich (Drescher 1992: 57):

> „Im Gegensatz zum bestimmten Artikel, der vor allem die Zugehörigkeit zur Gattung zum Ausdruck bringt, zeigt die Verwendung des unbestimmten Artikels im generischen Kontext an, daß es sich um eine ‚application particulière de la vérité générale‘ handelt […]. Während der generisch gebrauchte bestimmte Artikel die Zugehörigkeit des nachfolgenden Nomens zur Gattung herausstellt, funktioniert der generisch verwendete unbestimmte Artikel eher nach dem Modus der Exemplifizierung.“

Entsprechend kommt Drescher (ebd.: 59) zu dem Ergebnis, dass „[d] ie Nuancen in der Bedeutung der Artikel […] bei deren kommunikativer Verwendung offenbar nur eine untergeordnete Rolle“ spielen. Als wesentlichere Merkmale für die Beschreibung generalisierender Aussagen führt Drescher (ebd.: 64 ff.) das indefinite Personalpronomen *on* (ähnlich dem deutschen „man“) sowie das Tempus des Verbs an: Das Tempus Präsens sei die für generalisierende Aussagen typische Zeitform. Auf der Ebene des Lexikons identifiziert Drescher (ebd.: 68 f.) Frequenzadverbien wie *jamais* („nie“) und *toujours* („immer“) sowie Adverbien, die das Gewohnheitsmäßige einer Handlung oder eines Vorgangs zum Ausdruck bringen, z. B. *normalement* („normalerweise“) und *habituellement* („gewöhnlich“). Generell geht Drescher (ebd.: 70, Hervorhebungen im Original) davon aus, dass „[…] Abstrakta, die Wörter mit unspezifischer Intension und offener bis unbegrenzter Extension […] sowie

semantisch leere Verben wie *être* oder *avoir*“ für Aussagen mit Allgemeinheitscharakter typisch seien. Des Weiteren führt Drescher mit der generalisierenden Paraphrase, der generalisierenden Reihung und dem generalisierenden Vergleich auf Phrasen- oder Satzebene einzuordnende strukturelle Spielarten verallgemeinernder Aussagen an, die im Rahmen der hier unternommenen Untersuchung nicht operationalisiert werden können.
Auf Grundlage der Ausführungen Dreschers können die im letzten Abschnitt formulierten Forschungsfragen wie folgt ergänzt werden:

- Lexikalische Merkmale
    - Treten andere lexikalische Generalisierungssignale (z. B. Frequenzadverbien) häufig in wissenshaltigen Kontexten auf? Finden sich unpersönliche Pronomen?
    - Kann man Verben in *knowledge patterns* als semantisch „leere“ (generische) Verben bezeichnen?
- Strukturelle und grammatische Merkmale
    - In welcher grammatischen Form und in welchem Tempus ist das Definitionsprädikat realisiert? Finden sich unpersönliche Sätze, Passivkonstruktionen oder Präsentative?

5.2.1.3 Systemisch-funktionale Linguistik

Die SFL stellt einen elaborierten theoretischen Rahmen zur Beschreibung der Wechselbeziehungen zwischen linguistischen Merkmalsausprägungen an der Oberfläche von Sprachartefakten einerseits und ihrer konstitutiven Rolle für die Realisierung konkreter Sprachfunktionen (*metafunctions*) andererseits dar. Konkret versucht die SFL, einen theoretisch fundierten und für praktische Analysezwecke anwendbaren Rahmen für die Beschreibung der Interaktion zwischen linguistischen Metafunktionen und differenzierten Bedeutungsebenen sprachlicher Äußerungen bereitzustellen: „[…] meaning is construed by configurations of certain particular functions“ (Halliday 1994[2]: 34).

Die SFL erscheint aus mehreren Gründen als geeignetes Theoriegerüst für die hier angestrebte Ergänzung und Spezifizierung der Forschungsfragen für die linguistische Beschreibung wissenshaltiger Kontexte: Vorrangig ist in diesem Zusammenhang der Fokus der SFL auf die funktional-linguistische Analyse von Texten (vgl. ebd.: XV) zu nennen – entsprechend nennt Halliday (ebd.: XXIX) unter anderem die Anwendungsbereiche Sprachausbildung, Übersetzerausbildung und Lexikographie und damit Themen, die direkt oder indirekt zum Thema dieser Arbeit in Beziehung stehen. Besonders geeignet scheint die SFL daher als theoretischer Rahmen auch für die hier angestrebte korpuslinguistische Analyse von Texten: Einerseits sind die Merkmale von Sätzen, welche die SFL beschreibt, mit modernen Annotationsmethoden auch für größere Textmengen mit vertretbarem Aufwand zu erheben. Andererseits erlaubt die funktionale Ausrichtung der SFL die Rückkopplung von Merkmalsbeschreibungen an funktionale Kategorien und damit letztlich an klassische theoretische Konstrukte zur Beschreibung von Textsorten und Texttypen. Ein Beispiel für eine korpuslinguistische Untersuchung anhand von SFL-Kategorien ist etwa die Arbeit von Teich (2003) zu vergleichbaren Textkorpora sowie Übersetzungskorpora des Englischen und Deutschen.

Die systemisch-funktionale Linguistik ist aufgrund der Fülle an theoretischen Aussagen und der Elaboriertheit des Systems hier nicht vollständig darstellbar. Statt eine vollständige Darstellung des Systems der SFL zu versuchen, konzentriert sich vorliegende Arbeit daher auf diejenigen Aspekte, die für das hier unternommene Analysevorhaben als besonders relevant erscheinen: Prozesstypen und Metafunktionen. Zunächst geht die systemisch-funktionale Linguistik von drei grundlegenden Ebenen der Analyse sprachlicher Strukturen aus, denen wiederum jeweils eine Sprachfunktion

(*metafunction*) zugeordnet ist. Die Gesamtheit der Metafunktionen beschreibt dann den Bedeutungsgehalt der untersuchten Einheiten (vgl. Halliday 1994[2]: 35). Zu beachten ist dabei, dass als Analyseeinheit nicht Wörter oder Wortgruppen, sondern (Teil-)Sätze (*clauses*) gewählt werden (ebd.: 19, Hervorhebungen im Original):

> „[… O]ur main attention will be on the higher units, and particularly on the CLAUSE. This is because the mode of interpretation adopted here is a functional one, in which the grammatical structure is being explained by reference to the meaning; and there is a general principle in language whereby it is the larger units that function more directly in the realization of higher-level patterns."

Sätze können dabei zunächst in ihrer Eigenschaft als Informationsträger betrachtet werden: „A clause has meaning as a message, a quantum of information […]" (ebd.: 34). Auf dieser Analyseebene wird untersucht, mit welchen Mitteln Sätze ihre *textuelle Metafunktion* realisieren. Die textuelle Metafunktion stellt sicher, dass sprachliche Äußerungen als kommunikative Ereignisse wahrgenommen werden können, und prägt diesen eine Relevanz- bzw. Informationsstruktur auf (ebd.: 36 f.). Eine weitere Analyseebene ist die der sozialen Interaktion über Sprache, d. h. die Ebene des Satzes als Träger interaktionsbezogener Merkmale: „A clause has meaning as an exchange, a transaction between speaker and listener […]" (ebd.: 34). Mit den Mitteln dieser Ebene realisieren sprachliche Äußerungen ihre *interpersonelle Metafunktion*. Die letzte Analyseebene ist schließlich die der Darstellung außersprachlicher Sachverhalte: „A clause has meaning as a representation, a construal of some process in ongoing human experience […]" (ebd.). Hiermit ist folglich die Ebene der Semantik erreicht, entsprechend betont Halliday (ebd.: 106): „Usually when people talk about what a word or sentence 'means', it is this kind of meaning they have in mind – meaning in the sense of content." Die auf dieser Ebene untersuchten sprachlichen Mittel werden der *experientiellen Metafunktion* zugeordnet.

Den drei Metafunktionen ordnet die SFL Systeme sprachlicher Merkmale zu, mit deren Hilfe die Metafunktionen realisiert werden können. Für die textuelle Metafunktion ist das wesentliche Merkmal die Thema-Rhema-Struktur von Sätzen. Das Thema ist dabei dasjenige Element, auf das sich die Proposition des Satzes bezieht: „The Theme is the element which serves as the point of departure of the message; it is that with which the clause is concerned" (ebd.: 37). Das Thema ist, wenn es mit dem Subjekt des Satzes übereinstimmt, unmarkiert (ebd.: 43), d. h. es verfügt über keine besondere Emphase. Andere Themen werden als markiert, d. h. als sprachlich-inhaltlich hervorgehoben, bezeichnet (ebd.: 44). Das am häufigsten vorkommende unmarkierte Thema in alltagssprachlichen Aussagesätzen sei das Personalpronomen für die erste Person Singular (ebd.: 44), Beispiele für markierte Themen seien dagegen modale Ausdrücke oder Konjunktionen in satzinitialer Position (ebd.: 49).

Das wichtigste sprachliche Merkmal für die Realisierung der interpersonellen Metafunktion ist der Modus (ebd.: 68). Der Modus eines Satzes wird mit Hilfe des sogenannten finiten Operators ausgedrückt, der auf das Subjekt bezogen ist und durch die Merkmale Tempus, Modus und Polarität (An- oder Abwesenheit von Negationselementen) charakterisiert wird (ebd.: 72, 75). Weitere Elemente, die den Modus eines Satzes im Sinne der SFL markieren, sind die modalen und kommentierenden Adjunkte (*mood adjuncts*, *comment adjuncts*, ebd.: 82 f.). In dieser Gruppe wird gemäß der SFL-Terminologie ein breites Spektrum von Responsiven, Modalpartikeln und Adverbien (insbesondere temporale, frequentative, durative und modifikative Adverbien) zusammengefasst (zur deutschen Terminologie vgl. Hoffmann 2009).

Die experientielle Metafunktion wird insbesondere mittels eines elaborierten Systems von Prozesstypen realisiert (ebd.: 106). Halliday (ebd.: 107) unterscheidet zunächst zwischen psychologischen (*mental processes*) und materiellen (*material processes*) sowie relationalen (*relational processes*) Prozessen, zwischen denen jeweils Zwischenkategorien liegen. Materielle Prozesse können dabei als *processes of doing* (ebd.: 109 f.), in denen ein obligatorisches Agens eine Handlung vollzieht, charakterisiert werden. Ein psychologischer Prozess liegt dagegen bei *clauses of feeling, thinking and perceiving* (ebd.: 114) vor, die sich insbesondere dadurch auszeichnen, dass der Träger des Prozesses in der Regel menschlich bzw. mit einem Bewusstsein ausgestattet sein muss (ebd.). Halliday (ebd.: 116) betont des Weiteren, dass zwischen materiellen und psychologischen Prozessen auch dahingehend ein Unterschied vorliege, als das unmarkierte Tempus Präsens im Englischen das *simple present* für psychologische Prozesse und das *present progressive* für materielle Prozesse sei. Relationale Prozesse schließlich bezeichnet Halliday (ebd.: 119) als *processes of being*. Diese Gruppe wiederum zerfällt in die Subkategorien der Merkmalszuweisungen (*intensive processes*), der Umstandsaussagen (*circumstancial processes*) und der Eigentumsaussagen (*possessive processes*). Alle Subgruppen treten in zwei Formen auf, und zwar der attributiven Form der Merkmals- oder Klassenzuweisung (*attributive*) und in der Form von Identitätsaussagen (*identifying*, ebd.: 199 f.).

Entsprechend den Ausführungen zur SFL können die Forschungsfragen zur Untersuchung linguistischer Eigenschaften von KRCs nun ein letztes Mal ergänzt werden. Dabei ergibt sich aus der experientiellen Metafunktion die Frage nach den Prozesstypen bzw. – allgemeiner ausgedrückt – die Frage nach den semantischen Klassen, denen die Verben der prädikativen *knowledge patterns* zugeordnet werden können. Aus der textuellen Metafunktion folgt die Frage nach den in wissenshaltigen Kontexten verwendeten Formen der Thematisierung. Schließlich mündet die interpersonelle Metafunktion in die Untersuchung des Modus der in den KRCs auftauchenden Verben. In diesem Zusammenhang sollen auch Hinweise auf die syntaktische Komplexität der untersuchten KRCs berücksichtigt werden: Auch wenn Komplexitätsangaben nicht unter den Begriff des Modus fallen, gehört doch die Komplexität einer Äußerung zu den die Interaktion mit den Mitteln der Sprache wesentlich beeinflussenden Faktoren.

- Lexikalische Merkmale
    - Welchen semantischen Klassen kann man die Verben in *knowledge patterns* zuordnen?
- Strukturelle und grammatische Merkmale
    - Wie wird das Definiendum thematisiert?
    - Wie können die im Goldstandard vorkommenden Verben hinsichtlich des Modus charakterisiert werden?
    - Welche syntaktische Komplexität weisen wissenshaltige Kontexte auf?

Werden nun die aus der Literatur zu Definitionen und wissenshaltigen Kontexten, zu linguistischen Verfahren der Verallgemeinerung sowie der SFL entnommen Fragestellungen zusammengeführt, entsteht im Ergebnis dieser auf mehreren Ansätzen aufbauenden Vorgehensweise ein Katalog von Fragestellungen, deren Untersuchung zur Schärfung des stilistisch-linguistischen Profils wissenshaltiger Kontexte beitragen soll. Zwar lassen sich aus der SFL keine direkten Hypothesen zur linguistischen Realisierung wissenshaltiger Kontexte ableiten, die Ergebnisse der Untersuchung der entsprechenden Fragestellungen könnten allerdings Ansatzpunkte für weitere Studien auf neuem Datenmaterial bieten und die in der Literatur formulierten Hypothesen ergänzen. Tabelle 39 fasst alle formulierten Fragestellungen zusammen.

Ansatz/Quelle	Fragestellung
	Lexikalische Merkmale
Literatur zu wissenshaltigen Kontexten und Definitionen	Welche Wortarten treten häufig bzw. selten auf?
Literatur zu wissenshaltigen Kontexten und Definitionen	Welche lexikalischen Einheiten (Boostwörter) treten besonders häufig auf, finden sich ebenfalls generische Nomina, definitionstypische Adverbien oder andere lexikalische Generalisierungssignale?
Literatur zu linguistischen Verfahren der Verallgemeinerung	Treten andere lexikalische Generalisierungssignale (z. B. Frequenzadverbien) häufig in wissenshaltigen Kontexten auf? Finden sich unpersönliche Pronomen?
Literatur zu wissenshaltigen Kontexten und Definitionen	Welche lexikalischen Einheiten (Stopwörter) treten besonders selten auf?
Literatur zu wissenshaltigen Kontexten und Definitionen	Welche lexikalischen Einheiten kommen als *knowledge pattern* infrage?
Literatur zu linguistischen Verfahren der Verallgemeinerung	Kann man Verben in *knowledge patterns* als semantisch „leere“ (generische) Verben bezeichnen?
SFL	Welchen semantischen Klassen kann man die Verben in *knowledge patterns* zuordnen?
	Strukturelle und grammatische Merkmale
Literatur zu wissenshaltigen Kontexten und Definitionen, SFL	In welcher grammatischen und syntaktischen Form ist das Definiendum realisiert? Wie wird das Definiendum thematisiert?
SFL	Welche syntaktische Komplexität weisen wissenshaltige Kontexte auf?
Literatur zu linguistischen Verfahren der Verallgemeinerung	In welcher grammatischen Form und in welchem Tempus ist das Definitionsprädikat realisiert?
Literatur zu linguistischen Verfahren der Verallgemeinerung	Finden sich unpersönliche Sätze, Passivkonstruktionen oder Präsentative?
SFL	Wie können die im Goldstandard vorkommenden Verben hinsichtlich des Modus charakterisiert werden?
Literatur zu wissenshaltigen Kontexten und Definitionen	Finden sich Geltungsrestriktionen auf phraseologischer Ebene?
Literatur zu wissenshaltigen Kontexten und Definitionen	Wie lang sind wissenshaltige Kontexte?

Tabelle 39: Merkmalskatalog für die Untersuchung wissenshaltiger Kontexte

5.2.2 Lexikalische Merkmale

5.2.2.1 Methode

Zur Untersuchung lexikalischer Merkmale wissenshaltiger Kontexte werden die in Kapitel 4 erstellten KRC-Goldstandard-Datensätze wiederum mit Referenzdatensätzen verglichen, wobei Entwicklungs- und Testkorpora in ihrer Gesamtheit betrachtet, also nicht voneinander unterschieden werden, da weniger Eigenschaften der unterschiedlichen Fachgebiete von Interesse sind, als vielmehr linguistische Merkmale wissenshaltiger Kontexte insgesamt. Konkret werden als Referenzkorpora wie bereits in Kapitel 4 Unterabschnitte des RIC (für das Russische) und des NCL (für das Deutsche)

herangezogen. Den 632 deutschen KRCs werden damit 632 Sätze aus dem NCL gegenübergestellt, für das Russische umfassen Gold- und Vergleichsdaten jeweils 560 Sätze. Der Vergleich der Goldstandard-Daten mit diesen Sätzen soll Informationen über allgemein fachsprachliche bzw. für eher technische Texte spezifische – die Extraktionskorpora befassen sich mit als im weitesten Sinne „technisch“ zu charakterisierenden Fragestellungen – linguistische Eigenschaften erbringen. Da es sich bei den Sätzen des Goldstandards um aus ihrem textuellen Kontext herausgelöste Elemente handelt, scheint es erstrebenswert, dass auch die zum Vergleich herangezogenen Abschnitte der Referenzkorpora keine zusammenhängenden Texte bilden. Aus dem RIC wird dafür mit Hilfe eines Perl-Skripts eine entsprechende Zufallsstichprobe gezogen. Bei dem für das Deutsche herangezogene NCL scheint dagegen bereits eine Randomisierung stattgefunden zu haben, so dass die ersten 632 Sätze des Korpus‘ für die Untersuchung genutzt werden. Allerdings ist zusätzlich zu der Frage, inwieweit sich die annotierten wissenshaltigen Kontexte von zufällig ausgewählten Sätzen aus Referenzdaten unterscheiden, insbesondere die Frage von Interesse, welche diskriminatorische Kraft entsprechende Merkmale in Bezug auf die jeweiligen Extraktionskorpora selbst haben. Um Indizien für die Beantwortung dieser Frage zu finden, wird auch ein Vergleich der KRC-Goldstandard-Sätze mit einer äquivalenten Anzahl zufällig ausgewählter generischer Sätze (Nicht-KRCs) aus den in Abschnitt 4.3 beschriebenen Extraktionskorpora (Sub-EKR, Sub-TKR, Sub-EKD, Sub-TKD) durchgeführt.

Um solche lexikalischen Einheiten zu identifizieren, die in KRCs überproportional häufig (Boostwörter) bzw. selten (Stopwörter) auftreten, werden zunächst mit Hilfe eines Perl-Skripts nicht-alphabetische Zeichen aus den Datensätzen entfernt. Alsdann werden wiederum mit Hilfe von TreeTagger (Schmid 1994) Wortartenannotation und Lemmatisierung durchgeführt[91]. Für das Russische wird dabei im Unterschied zur bisherigen Darstellung mit dem *tagset* „Russian small“[92] ein *tagset* verwendet, das auf der Wortformenannotation des Russischen Nationalkorpus‘[93] beruht, da die feingliedrige Annotation mit Hilfe des bisher verwendeten *tagsets* für die hier beschriebene Untersuchung unerwünschte Unterscheidungen (feingliedrige morphologische Analyse) trifft. Anhand der getaggten und lemmatisierten Datensätze wird alsdann ein Vergleich der Häufigkeiten von Wortarten sowie von Lemmata durchgeführt. Hierfür wird dieselbe Methode verwendet wie für den in Abschnitt 4.1.4.1 beschriebenen Vergleich der Babouk-Korpora mit Referenzkorpora (Chi-Quadrat- bzw. Fisher-Test und Konfidenzintervalle).

Für die Identifizierung und Beschreibung möglicher Bausteine für *knowledge patterns* soll lediglich ein Teil des erstellten russischen und deutschen Goldstandards untersucht werden, und zwar derjenige Teil, der für die entsprechenden Entwicklungskorpora annotiert wurde (Sub-EKD und Sub-EKR). Auf diese Weise wird sichergestellt, dass für die später durchzuführende Evaluierung einer automatischen Extraktionsmethode auf den erstellten Testkorpora (Sub-TKD und Sub-TKR) ungesehenes Material verwendet werden kann und eine Feinabstimmung von Extraktionsmustern auf die Testkorpora vermieden wird, so dass die Möglichkeit einer Überprüfung der Übertragbarkeit sprachlicher Muster auf neue Korpora und Fachgebiete gewahrt bleibt. Für die Interpretation und Bewertung von Ergebnissen für solche Einheiten, für die auch Ergebnisse aus der Untersuchung von Frequenzen im gesamten Goldstandard im Vergleich zu einer äquivalenten Satzmenge aus einem Referenzkorpus vorliegen, wird indes auch auf statistische Vergleichsdaten für den gesamten Goldstandard (d. h. aus den jeweiligen Entwicklungs- und Testkorpora gemeinsam) zurückgegriffen,

[91] Für die Analyse der RIC-Sätze wird die bereits bestehende Lemmatisierung genutzt.

[92] http://corpus.leeds.ac.uk/mocky/, abgerufen am 08.04.2014.

[93] Das Korpus ist online unter http://www.ruscorpora.ru/en/index.html einsehbar, eine Beschreibung des Tagsets findet sich hier: http://www.ruscorpora.ru/download/tagset_rus.pdf, abgerufen am 08.04.2014.

so dass Aussagen darüber getroffen werden können, ob im Hinblick auf die in den untersuchten Sätzen identifizierten Elemente Verteilungsunterschiede zwischen dem gesamten Goldstandard und dem jeweiligen Vergleichskorpus vorliegen.
Für die Untersuchung potentieller Bausteine von *knowledge patterns* werden in den Goldstandard-KRCs der Entwicklungskorpora manuell diejenigen lexikalischen und nicht-lexikalischen Elemente identifiziert, die Bestandteile von Begriffserläuterungen im Sinne der Definition wissenshaltiger Kontexte einleiten bzw. diese sprachlich markieren und einen direkten linguistischen Bezug zum jeweiligen Definiendum aufweisen. In der Regel handelt es sich hierbei um einfache Prädikate (ebenfalls zum Prädikat zählende Modal- und Hilfsverben werden in der Darstellung nicht berücksichtigt), allerdings werden auch lexikalische Elemente mit anderer syntaktischer Funktion und nicht-lexikalische Elemente wie beispielsweise Parenthesen dann als Bausteine für *knowledge patterns* akzeptiert, wenn sie für die Formulierung von Begriffserläuterungen eine wesentliche Rolle spielen, d. h. insbesondere dann, wenn sie das Vorliegen einer solchen Erläuterung signalisieren bzw. diese direkt einleiten. Sofern in den untersuchten Sätzen mehr als ein Prädikat bzw. mehr als ein Kandidat für einen Musterbestandteil auftauchen, werden in den Sätzen mehrere Bausteine für *knowledge patterns* identifiziert. Um eine möglichst große Genauigkeit bei der Ermittlung der relevanten lexikalischen Einheiten zu erreichen, werden Letztere in ihrer grammatischen Grundform dargestellt, was bedeutet, das auch von Verben abgeleitete Adjektive (z. B. *установить*, „aufstellen“ –> *установленный*, „der aufgestellte“) als Verben dargestellt werden. Gerade für das Russische wird durch diese manuelle „Lemmatisierung“ eine höhere Präzision bei der Aspektzuordnung der identifizierten Verben sowie letztlich der Ermittlung der Kandidaten für die Bildung von *knowledge patterns* erreicht als bei automatischer Lemmatisierung.

5.2.2.2 Wortartenfrequenzen

5.2.2.2.1 Russisch

Abbildung 14 stellt die Ergebnisse des Vergleichs der Wortartenfrequenzen in den jeweils 560 KRCs des Goldstandards, der RIC-Stichprobe und der Stichprobe aus den Extraktionskorpora Sub-EKR und Sub-TKR graphisch dar: Konkret handelt es sich um eine Darstellung der relativen Frequenzen der für den Goldstandard 11 häufigsten Wortarten, die nach Entfernung nicht-alphabetischer Zeichen im Korpus verbleiben.
Der Vergleich der Wortartenfrequenzen zwischen den KRCs des russischen Goldstandards und einem Vergleichsdatensatz aus dem RIC bestätigt zunächst die Darstellung aus Tabelle 8 (Abschnitt 4.1.4.1) zur Häufigkeit von Substantiven (*tag* S) und Adverbien (*tag* ADV), mit deren Hilfe eine tentative Fachlichkeitsbewertung der gesammelten Web-Korpora vorgenommen wurde. Zu dem bereits dort dargestellten Bild passt auch, dass der Goldstandard eine deutlich höhere Frequenz von ebenfalls zur Nominalgruppe zählenden Wortarten – insbesondere Adjektiven (*tag* A) – aufweist als die Vergleichsdaten. Im Vergleich zu den KRCs des Goldstandards zeigen die Vergleichssätze des RIC allerdings – zusätzlich zur bisherigen Darstellung in Bezug auf Nomina und Adverbien – einen höheren Anteil an Verbformen (*tag* V), Konjunktionen (*tag* CONJ), Pronomen (*tags* A-PRO, S-PRO, ADV-PRO) und Partikeln (*tag* PART). Geringere Unterschiede scheinen dagegen hinsichtlich der Anteile an Präpositionen (*tag* PR) und Numeralia (*tag* NUM) in den beiden Datensätzen zu bestehen. Im Vergleich zu der Stichprobe aus den Extraktionskorpora Sub-EKR und Sub-TKR ergeben sich für die

KRCs des Goldstandards – geringere – Unterschiede in Bezug auf Nominalgruppen, Verben, Adverbien und Partikel.

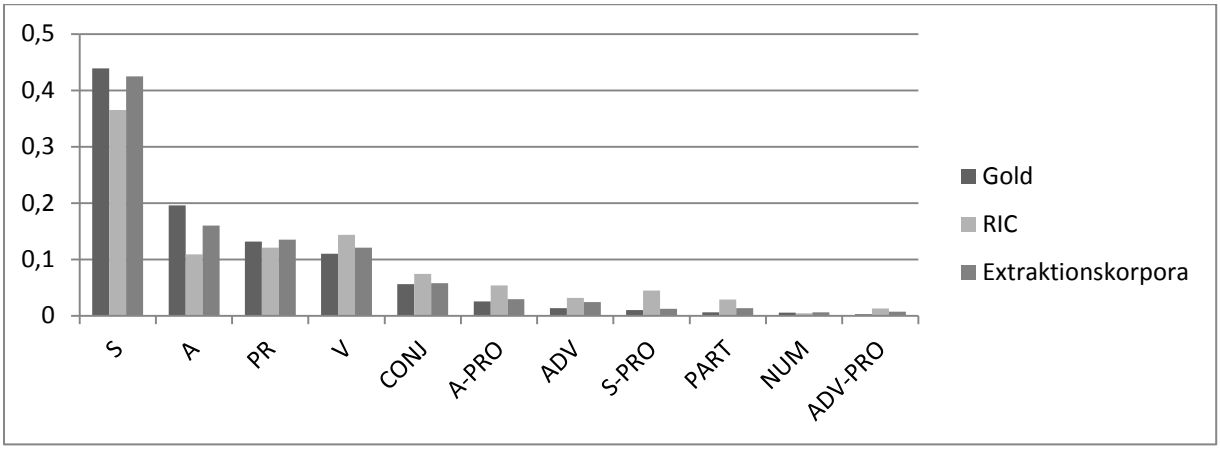


Abbildung 14: Relative Wortartenfrequenzen im russischen KRC-Goldstandard, in einer Stichprobe aus dem RIC sowie aus den Extraktionskorpora

Tabelle 40 stellt im Ergebnis der Berechnung die absoluten und relativen Frequenzen aller *tags* in den untersuchten Datensätzen sowie die Prüfgröße des Chi-Quadrat-Tests, den p-Wert und das 95%-Konfidenzintervall für den Unterschied zwischen den miteinander verglichenen Gold- bzw. RIC-Anteilswerten (alle Rundungen erfassen vier Nachkommastellen) zusammen[94]. *n* bezeichnet dabei die Gesamtzahl der untersuchten Wörter in den jeweiligen Datensätzen. Fett hervorgehoben sind diejenigen Zeilen, in denen eine bestimmte Wortart relativ häufiger im Goldstandard als im Vergleichskorpus erscheint. Kursive Markierung weist dagegen darauf hin, dass die betreffende Wortart im Goldstandard relativ seltener erscheint als in den Vergleichsdaten. Fälle, in denen der Fisher-Test angewandt wurde, sind durch den niedriggestellten Buchstaben F ($_F$) ausgezeichnet. Zeilen mit einem auf einem 5%-Signifikanzniveau nicht-signifikanten Testergebnis werden zudem abgegraut dargestellt. Merkmale, für die auch beim Vergleich der KRC-Goldstandard-Sätze mit generischen Sätzen aus den Extraktionskorpora ein signifikantes Ergebnis ermittelt wurde, sind mit * markiert, in Klammern sind dann die entsprechende Prüfgröße und der p-Wert angegeben.

Tag	Absolute Frequenz Gold-standard (n=10575)	Relative Frequenz Gold-standard	Absolute Frequenz RIC (n=9378)	Relative Frequenz RIC	Prüfgröße	P	KI (95%)
S*	**4643**	**0,4391**	**3425**	**0,3652**	**112,2040 (4,1150)**	**< 0,0001 (0,0425)**	**[0,0602, 0,0875]**
A*	**2068**	**0,1956**	**1023**	**0,1091**	**283,2143 (40,7758)**	**< 0,0001 (< 0,0001)**	**[0,0765, 0,0964]**
PR	**1390**	**0,1314**	**1133**	**0,1208**	**4,9863**	**0,0256**	**[0,0013, 0,0199]**

[94] Da es sich bei dem Fisher-Test um einen exakten Test handelt, bei dem das Testergebnis nicht mit Hilfe einer Prüfgröße angenähert werden muss, wird in Fällen, in denen dieser Test angewandt wurde, nur der p-Wert angegeben. Das angegebene Konfidenzintervall bezieht sich dann auf das Chancen-Verhältnis (vgl. Gries 2009: 174).

*V**	*1162*	*0,1099*	*1345*	*0,1434*	*50,5870 (5,7792)*	*< 0,0001 (0,0162)*	*[-0,0429, -0,0242]*
CONJ	*592*	*0,0560*	*695*	*0,0741*	*26,7708*	*< 0,0001*	*[-0,0251, -0,0112]*
A-PRO	*269*	*0,0254*	*504*	*0,0537*	*106,1740*	*< 0,0001*	*[-0,0339, -0,0227]*
*ADV**	*142*	*0,0134*	*298*	*0,0318*	*76,7455 (31,4932)*	*< 0,0001 (< 0,0001)*	*[-0,0226, -0,0141]*
S-PRO	*109*	*0,0103*	*422*	*0,0450*	*229,5814*	*< 0,0001*	*[-0,0394, -0,0300]*
*PART**	*65*	*0,0061*	*273*	*0,0291*	*156,0198 (25,8465)*	*< 0,0001 (< 0,0001)*	*[-0,0268, -0,0191]*
NUM	**61**	**0,0058**	**39**	**0,0042**	**2,2698**	**0,1319**	**[-0,0004, -0,0037]**
*ADV-PRO**	*32*	*0,0030*	*120*	*0,0128*	*61,4693 (16,7911)*	*< 0,0001 (< 0,0001)*	*[-0,0124, -0,0072]*
A-NUM	*15*	*0,0014*	*22*	*0,0023*	*1,8360*	*0,1754*	*[-0,0022, 0,0004]*
INIT	*13*	*0,0012*	*16*	*0,0017*	*0,4847*	*0,4863*	*[-0,0016, 0,0007]*
*PRAEDIC**	*11*	*0,0010*	*60*	*0,0064*	*38,7422 (17,7076)*	*< 0,0001 (< 0,0001)*	*[-0,0072, -0,0035]*
$INTJ_F$	3	0,0003	3	0,0003	- (4,8065)	1 (0,0284)	[0,1187, 6,6228]

Tabelle 40: Unterschiede in den Wortartenfrequenzen des russischen Goldstandards, des RIC und der Extraktionskorpora

Im Ergebnis zeigt sich, dass neben Elementen von Nominalphrasen (*tags* S, A) auch Verben, Präpositionen und Adverbien (*tags* V, PR und ADV), Konjunktionen (*tag* CONJ), Pronomina (*tags* A-PRO, S-PRO, ADV-PRO), Partikel (*tag* PART) und Prädikative (*tag* PRAEDIC) sehr signifikante Verteilungsunterschiede zwischen dem russischen Goldstandard und der RIC-Stichprobe aufweisen. Der statistische Vergleich der Wortartenverteilungen bestätigt damit zunächst die bereits oben formulierten Aussagen. Im Vergleich des KRC-Goldstandards mit Nicht-KRCs aus denselben Korpora (Sub-EKR und Sub-TKR) können – durchgängig deutlich kleinere – Verteilungsunterschiede allerdings nur für Substantive, Adjektive, Adverbien, Pronomen (ADV-PRO), Verben, Partikel und Prädikative bestätigt werden, andererseits zeigt sich auch ein bislang nicht beobachteter Unterschied für den *tag* INTJ (Interjektionen, Prüfgröße 4,8065, p = 0,0284, wobei der Anteil *a* im Goldstandard geringer ist als in den Referenzdaten: $a_G < a_R$). Der höhere Anteil an Nominalgruppen im Goldstandard im Vergleich zur RIC-Stichprobe kann möglicherweise auf den höheren Anteil an Termini und Nominalisierungen zurückgeführt werden und entspricht einem bereits diskutierten quantitativen Merkmal, das Fachtexten zugeschrieben wird (vgl. Abschnitt 4.1.4.1). Im Vergleich zu Stichproben aus den Extraktionskorpora Sub-EKR und Sub-TKR weist dieser Unterschied jedoch möglicherweise eher auf die explizite Benennung relevanter Begriffe mit Hilfe von Termini hin, womit einem der in Kapitel 4 formulierten Selektionskriterien entsprochen wäre. Der geringere Anteil an Verben, Adverbien und Prädikativen steht möglicherweise mit einem reduzierten Stil, d. h. einem tendenziellen Verzicht auf modale Abtönung von Aussagen durch Hilfsverben, Adverbien und Prädikative wie *можно* (man kann), *пора* (es ist Zeit), *жаль* (schade), in Beziehung. Das seltenere Auftreten von Pronomina kann in diesem Zusammenhang möglicherweise ebenfalls als Resultat eines funktionalen Explizitheitsgebots interpretiert werden und entspricht damit einem der Selektionskriterien bei der

Erstellung des Goldstandards, allerdings deuten die Ergebnisse der Untersuchung darauf hin, dass es sich hierbei um ein generelles Merkmal der eher technischen Texte aus den Babouk-Korpora im Vergleich zum RIC handelt. Die Verteilungsunterschiede für den *tag* ADV-PRO, mit dem Lemmata wie *как* („wie“), *так* („so“), *где* („wo“), *иначе* („anders“) und *здесь* („hier“) markiert werden, weisen dagegen eher darauf hin, dass in russischen KRCs Einheiten wie etwa Deixis und Fragewörter seltener verwendet werden als in den Vergleichsdaten, was mit den Aussagen Walters (2010) in Bezug auf Stopwörter für das Deutsche übereinstimmt. Das seltenere Auftreten von Partikeln einschließlich des Modalpartikels passt zu einer funktionalen Sicht auf wissenshaltige Kontexte, da beispielsweise Negations- und Irrealispartikel mit den funktionalen Anforderungen der expliziten Darstellung bestehender Sachverhalte nur schwer in Einklang zu bringen sind, und entspricht ebenfalls einem der in der Literatur beschriebenen Merkmale. Hinzuweisen ist in diesem Zusammenhang jedoch wiederum darauf, dass die automatische Wortartenannotation nicht fehlerfrei funktioniert. Angesichts der sehr geringen Anteile einiger Wortarten an den Datensätzen ist zudem davon auszugehen, dass gesicherte Aussagen nur aufgrund umfangreicherer Untersuchungen möglich sind. Eine systematische Untersuchung der Frequenzen von Wortarten in wissenshaltigen Kontexten und anderen Satztypen unter Berücksichtigung von Annotationsfehlern und auf Grundlage umfangreicherer Datensätze könnte Gegenstand eingehenderer Untersuchungen sein. Die beschriebenen Unterschiede zwischen den Extraktionskorpora und den Gold-Daten müssen jedoch aufgrund ihrer geringen Größe – einzig die Gruppe der Adjektive erreicht mit dem Konfidenzintervall KI = [0,0245, 0,0461] für den Unterschied zwischen den beiden Anteilswerten eine relevante Stärke – als eher marginale und potentiell unzuverlässige Merkmale eingestuft werden.

5.2.2.2.2 Deutsch

Analog zur Darstellung für das Russische visualisiert Abbildung 15 die relativen Frequenzen der zehn häufigsten Wortarten des deutschen KRC-Goldstandards im Vergleich zu einer Stichprobe aus dem NCL sowie einer Zufallsstichprobe aus den deutschen Extraktionskorpora Sub-EKD und Sub-TKD. Ähnlich wie bereits für das Russische und wiederum in Übereinstimmung mit Abschnitt 4.1.4.1 fällt auch hier die deutlich größere Häufigkeit von Nominalgruppen – Substantiven (NN), Artikeln (ART) und attributiv gebrauchten Adjektiven (ADJA) – im Goldstandard im Vergleich zum NCL ins Auge und auch hier weisen beide Vergleichskorpora einen höheren Anteil von Adverbien (ADV) auf als die KRCs des Goldstandards. Darüber hinaus enthält der Goldstandard einen höheren Prozentsatz an Präpositionen (APPR) als das NCL, wobei der Vergleich mit dem Extraktionskorpus diese Tendenz aber nicht bestätigt. Auffällig ist zudem der höhere Prozentsatz finiter Verben (VVFIN) im NCL, wobei jedoch auch diese Tendenz durch das Extraktionskorpus nicht bestätigt wird. Geringfügige Unterschiede zeigen sich in Bezug auf die *tag*s VVPP (Partizip II) und ADJD (prädikativ gebrauchte Adjektive). Annähernd gleich verteilt scheinen dagegen finite Hilfsverben (VAFIN) zu sein. Auch hier bieten sich eingehendere Untersuchungen an. Nebenordnende Konjunktionen (KON) erscheinen im Goldstandard etwas häufiger als als im NCL und als in den Extraktionskorpora. Wie bereits für das Russische fallen die Unterschiede zwischen den Gold-Daten und den Extraktionskorpora generell deutlich geringer aus als diejenigen zwischen den Gold-Daten und dem NCL.

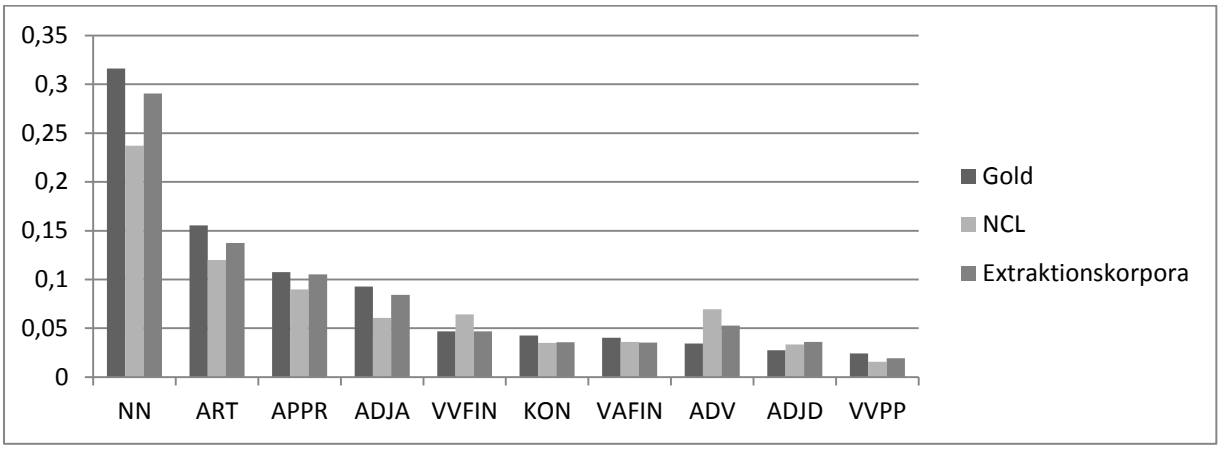


Abbildung 15: Relative Wortartenfrequenzen im deutschen KRC-Goldstandard, in einer Stichprobe aus dem NCL sowie aus den Extraktionskorpora

Tabelle 41 fasst analog zu Tabelle 40 die statistischen Daten nur für diejenigen Wortarten zusammen, für die mit Hilfe der durchgeführten Tests signifikante Verteilungsunterschiede ermittelt werden konnten: Da das STTS[95] wesentlich mehr *tags* umfasst als „Russian small" (das hier beschriebene Experiment umfasste 44 *tags*), wäre eine vollständige Darstellung für die hier verfolgten Zwecke zu umfangreich. Wiederum durch * markiert werden die Verteilungsunterschiede, die auch durch den Vergleich der KRC-Goldstandard-Daten mit den Zufallsstichproben aus den Extraktionskorpora Sub-EKD und Sub-TKD bestätigt werden können, die Ergebnisse des entsprechenden Signifikanztests sind wieder in Klammern angegeben.

Tag	Absolute Frequenz Gold-standard (n=12073)	Relative Frequenz Gold-standard	Absolute Frequenz NCL (n=9265)	Relative Fre-quenz NCL	Prüfgröße	p	KI (95%)
NN*	**3817**	**0,3162**	**2197**	**0,2371**	**161,3709 (18,0604)**	**< 0,0001 (< 0,0001)**	**[0,0669, 0,0911]**
ART*	**1877**	**0,1555**	**1111**	**0,1199**	**54,7420 (15,0738)**	**< 0,0001 (0,0001)**	**[0,0262, 0,0449]**
APPR	**1298**	**0,1075**	**831**	**0,0897**	**18,3358**	**< 0,0001**	**[0,0097, 0,0259]**
ADJA*	**1122**	**0,0929**	**563**	**0,0608**	**74,1418 (5,2046)**	**< 0,0001 (0,0225)**	**[0,0250, 0,0394]**
VVFIN	*567*	*0,0470*	*597*	*0,0644*	*30,6893*	*< 0,0001*	*[-0,0238, -0,0111]*
KON*	**516**	**0,0427**	**325**	**0,0351**	**7,9269 (7,3773)**	**0,0049 (0,0066)**	**[0,0024, 0,0130]**
*ADV**	*417*	*0,0345*	*644*	*0,0695*	*134,9242 (46,7782)*	*< 0,0001 (< 0,0001)*	*[-0,0412, -0,0288]*
*ADJD**	*331*	*0,0274*	*309*	*0,0334*	*6,1439 (13,6784)*	*0,0132 (0,0002)*	*[-0,0107, -0,0012]*
VVPP*	**293**	**0,0243**	**145**	**0,0157**	**18,9415 (6,9852)**	**< 0,0001 (0,0082)**	**[0,0048, 0,0124]**

[95] Stuttgart-Tübingen-Tagset. Eine Übersicht über die verwendeten *tags* findet sich hier: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, abgerufen am 08.04.2014.

*NE**	*105*	*0,0087*	*492*	*0,0531*	*378,4649 (107,2427)*	*< 0,0001 (< 0,0001)*	*[-0,0494, -0,0395]*
PIS	*85*	*0,0070*	*103*	*0,0111*	*9,5143*	*0,0020*	*[-0,0068, -0,0014]*
*PIAT**	*68*	*0,0056*	*110*	*0,0119*	*23,9278 (6,0755)*	*< 0,0001 (0,0137)*	*[-0,0089, -0,0036]*
*VMFIN**	*67*	*0,0055*	*132*	*0,0142*	*41,9853 (12,0008)*	*< 0,0001 (0,0005)*	*[-0,0115, -0,0058]*
*VVINF**	*65*	*0,0054*	*116*	*0,0125*	*30,8989 (5,5164)*	*< 0,0001 (0,0188)*	*[-0,0098, -0,0044]*
PRF	*63*	*0,0052*	*79*	*0,0085*	*8,1868*	*0,0042*	*[-0,0057, -0,0009]*
*PPER**	*60*	*0,0050*	*299*	*0,0323*	*234,5802 (29,8524)*	*< 0,0001 (< 0,0001)*	*[-0,0312, -0,0234]*
PTKZU	*39*	*0,0032*	*53*	*0,0057*	*7,0026*	*0,0081*	*[-0,0044, -0,0006]*
PDS	*38*	*0,0031*	*59*	*0,0064*	*11,3138*	*0,0008*	*[-0,0052, -0,0012]*
PTKVZ	*31*	*0,0026*	*41*	*0,0044*	*4,8405*	*0,0278*	*[-0,0036, -0,0001]*
*PTK-NEG**	*31*	*0,0026*	*87*	*0,0094*	*43,1361 (8,0207)*	*< 0,0001 (0,0046)*	*[-0,0091, -0,0046]*
*KOUS**	*26*	*0,0022*	*67*	*0,0072*	*29,9903 (31,2354)*	*< 0,0001 (< 0,0001)*	*[-0,0071, -0,0031]*
*PPO-SAT**	*24*	*0,0020*	*94*	*0,0101*	*61,9609 (9,2802)*	*< 0,0001 (0,0023)*	*[-0,0104, -0,0059]*
VAPP	*4*	*0,0003*	*12*	*0,0013*	*5,2772*	*0,0216*	*[-0,0019, -0,0001]*
PWS_F	*1*	*0,0001*	*8*	*0,0009*	*-*	*0,0129*	*[0,0022, 0,7152]*

Tabelle 41: Unterschiede in den Wortartenfrequenzen des deutschen Goldstandards, des NCL und der Extraktionskorpora

Ebenso wie bereits für das Russische wird deutlich, dass sich der Goldstandard im Vergleich zum NCL durch einen geringeren Anteil an Pronomen (*tags* PDS, PIS, PIAT, PRF, PPER, PPOSAT, PWS), Partikeln (insbesondere von Negationspartikeln: PTKNEG, aber auch abgetrennter Verbpartikel: PTKVZ und des „zu"-Partikels vor einem Infinitiv: PTKZU) und bestimmten Verbformen (VMFIN: finites Modalverb, VVINF: Infinitiv, VVFIN: finite Vollverben) auszeichnet. Während nebenordnende Konjunktionen (*tag* KON) im Goldstandard häufiger vertreten sind als in den Vergleichsdaten, nehmen unterordnende Konjunktionen (*tag* KOUS) einen geringen Anteil ein. Des Weiteren bestätigt Tabelle 41 die Signifikanz der bereits im Zusammenhang mit Abbildung 15 hervorgehobenen Verteilungsunterschiede. Anders als das für das Russische verwendete *tagset* „Russian small" zeichnet das STTS auch Eigennamen (NE) aus, für die im Goldstandard deutlich weniger Belege gefunden werden als in den Vergleichsdaten. Gar nicht im Goldstandard, wohl aber im NCL vertreten sind Zirkumpositionen nach rechts (*tag* APZR), Interjektionen (*tag* ITJ), unterordnende Konjunktionen mit nachfolgendem Infinitiv mit „zu" (*tag* KOUI) sowie Antwortpartikel (*tag* PTKANT).

Die Analyse der deutschen Daten ermöglicht es, die für das Russische erarbeiteten Ergebnisse in Teilen zu bestätigen und zu präzisieren. Die ermittelten Verteilungsunterschiede im Vergleich zum NCL sind hoch signifikant. Die untersuchten KRCs zeichnen sich im Vergleich zu den 632 untersuchten Sätzen aus dem NCL durch einen höheren Anteil von Nominalgruppen (die *tags* NN,

ART und ADJA weisen jeweils relevante Effektstärken – gemessen anhand der jeweiligen Konfidenzintervalle – auf) aus, was wiederum als Indiz für Fachsprachlichkeit interpretiert werden kann. Ebenso wie bereits im Fall der russischen Daten zeigen die deutschen KRCs niedrigere Anteile an Pronomina (insbesondere irreflexiver Personalpronomina, *tag* PPER, die als einzige Gruppe der Pronomina allein eine deutlich über 1% liegende Effektstärke aufweisen[96]) und Partikeln. Auch die Gruppen der Adverbien (ADV) und der Eigennamen (NE) sind in den beiden Datensätzen stark unterschiedlich verteilt (hoch-signifikante Ergebnisse bei relevanter Effektstärke). Dieser Befund stützt wiederum die Annahme eines Funktionalstils von KRCs, der Explizitheit und damit den weitgehenden Verzicht auf Pronomina ebenso vorschreibt wie die Formulierung positiver Sachaussagen (Verzicht auf Negation, wobei die Negationspartikel PTKNEG aufgrund ihres geringen Anteils auch nur eine geringe Effektstärke aufweisen, so dass dieses Merkmal nur eine geringe diskriminatorische Kraft hat). Der geringe Anteil an Eigennamen (NE) im Goldstandard stützt die Annahme eines eher abstrakten, generalisierenden Stils mit nur wenigen Bezügen auf konkrete Einzelfälle. In Bezug auf Verbformen und Konjunktionen erlauben die deutschen Daten eine differenziertere Darstellung als die russischen Daten: So enthält der Goldstandard mindestens 1% weniger finite Verbformen (VVFIN) als der Vergleichsdatensatz aus dem NCL, wohingegen die Unterschiede in Bezug auf Modalverben (VMFIN), Infinitive (VVINF) und Partizipien (VVPP) allein schon wegen der geringen Häufigkeit dieser *tags* geringer ausfallen. Der Verzicht auf Modalverben und auch Adverbien deutet wie bereits für das Russische auf einen reduzierten, unpersönlichen Stil hin. Die überwiegende Verwendung nebenordnender Konjunktionen (*tag* KON im Gegensatz zu unterordnenden Konjunktionen, *tag* KOUS) deutet darüber hinaus auf ein Überwiegen parataktischer Satzstrukturen hin.

Die im Vergleich zum NCL ermittelten Unterschiede können beim Vergleich der KRC-Goldstandard-Sätze mit der Zufallsstichprobe aus den Extraktionskorpora Sub-EKD und Sub-TKD – wiederum bei geringeren Effektgrößen – für die *tags* NN, ART, ADJA, KON, VVPP, NE, PIAT, VMFIN, VVINF, PPER, PTKNEG, KOUS und PPOSAT bestätigt werden, zusätzlich ergibt dieser Vergleich Unterschiede für finite Hilfsverben (VAFIN, Prüfgröße 3,7510, p = 0,0528, $a_G > a_R$), Kardinalzahlen (CARD, Prüfgröße 5,7245, p = 0,0167, $a_G > a_R$), Pronominaladverbien („dafür“, „trotzdem“, *tag* PAV, Prüfgröße 11,2503, p = 0,0008, $a_G < a_R$) und attribuierende Demonstrativpronomina (PDAT, Prüfgröße 30,7665, p < 0,0001, $a_G < a_R$). Allerdings ist im Hinblick auf diese Ergebnisse auf die geringen anhand der Konfidenzintervalle geschätzten Effektstärken zu verweisen – lediglich für die *tags* NN, ADV und NE ergibt der Vergleich Effektstärken von mindestens 1% für den Unterschied der Anteilswerte in den KRC-Gold-Daten und den Extraktionskorpora Sub-EKD und Sub-TKD, so dass gesicherte Aussagen nur auf Grundlage umfangreicherer Untersuchungen zu machen sind, die hier ermittelten Ergebnisse aber lediglich Tendenzen bzw. eher marginale Merkmale der miteinander verglichenen Datensätze beschreiben.

[96] Werden alle Arten von Pronomina jedoch zusammengenommen, ergibt der statistische Test für das *setting* Goldstandard-NCL ein hoch signifikantes Ergebnis (p < 0,0001) bei einer Prüfgröße χ^2 = 303,4135. Das Konfidenzintervall *KI* beträgt *KI* = [-0,0595, -0,0467] und beschreibt damit eine relevante, d. h. Frequenzunterschiede im dreistelligen Bereich erreichende Effektgröße. Für die Gruppe der Indefinitpronomen (*tags* PIS und PIAT) ergibt sich eine Prüfgröße χ^2 = 32,4877 bei p < 0,0001 und ein Konfidenzintervall *KI* von *KI* =[-0,0141, -0,0066].

5.2.2.3 Boost- und Stopwörter

5.2.2.3.1 Russisch

In diesem Abschnitt soll untersucht werden, welche lexikalischen Einheiten im Goldstandard häufiger bzw. seltener auftreten als in Referenzdatensätzen. Als Stopwörter sollen dabei solche Einheiten bezeichnet werden, die für das RIC bzw. die „Babouk“-Stichproben charakteristisch sind und im Goldstandard nur selten oder gar nicht auftreten. Solche Einheiten können somit als für wissenshaltige Kontexte wenig charakteristisch bezeichnet werden. Als Boostwörter werden im Gegensatz dazu lexikalische Einheiten bezeichnet, die im Goldstandard mit einem höheren Anteil vertreten sind als in den Vergleichsdaten und somit für wissenshaltige Kontexte typisch zu sein scheinen. Die Untersuchung strebt dabei an, möglichst allgemeingültige, d. h. nicht nur für die hier konkret untersuchten Datensätze, sondern auch darüber hinaus geltende Aussagen zu treffen, indem sie eindeutig terminologische oder textspezifische, z. B. durch häufige Wiederholungen ähnlicher oder sehr ähnlicher Aussagen gehäuft auftretende Einheiten, ausschließt, d. h. im Allgemeinen solche Einheiten, für die ohnehin mit Verteilungsunterschieden zwischen dem Gold- und dem Referenzkorpus gerechnet werden muss. Für die Untersuchung statistischer Verteilungsunterschiede wird dasselbe Verfahren wie in den vorigen Abschnitten verwendet. Hinzuweisen ist in diesem Zusammenhang allerdings darauf, dass ein solches statistisches Verfahren eine Auswahl funktionalstilistisch motivierter, d. h. nicht-terminologischer oder textspezifischer Verteilungsunterschiede, nicht leisten kann, so dass diese eher als qualitativer Prozess – etwa bei der Bewertung des terminologischen Status‘ eines Lemmas – zu charakterisieren ist. Beispielsweise werden solche Lemmata nicht berücksichtigt, die zwar auch als generische Nomina auftreten, jedoch überwiegend als Bestandteil terminologischer Einheiten im Goldstandard erscheinen (beispielsweise *устройство,* „Anlage“, „Apparat“), oder aber solche Lemmata, die für das – zugegebenermaßen hier nur sehr grob umrissene – Fachgebiet Technik, aus dem die KRCs des Goldstandards mehrheitlich stammen, eher typisch scheinen (z. B. *корпус*, „Korpus“, „Gehäuse“). Tabelle 42 fasst die Ergebnisse für russische Lemmata, für die signifikante Verteilungsunterschiede ermittelt werden konnten und die möglicherweise Indikatoren für den angenommenen Funktionalstil wissenshaltiger Kontexte darstellen, also nicht als akzidentell oder terminologisch angesehen werden, in der hier gebotenen gedrängten Form zusammen und ordnet die einzelnen Einheiten linguistischen Kategorien zu. Einheiten, für welche Verteilungsunterschiede auch im Hinblick auf die Extraktionskorpora Sub-EKR und Sub-TKR ermittelt werden konnten, werden wiederum mit * markiert, in Klammern werden wiederum die Ergebnisse des entsprechenden Signifikanztests vermerkt. Die Ergebnisse werden alsdann anhand einiger Beispiele diskutiert, wobei die fraglichen lexikalischen Einheiten zum besseren Verständnis durch Unterstreichung hervorgehoben sind. In den deutschen Übersetzungen der russischen Beispiele finden sich diese Einheiten nicht immer wortgetreu wieder, allerdings sind auch in den deutschen Übersetzungen diejenigen Einheiten, die am ehesten den diskutierten russischen Lexemen entsprechen, durch Unterstreichung markiert.

Wort	Übersetzung	Absolute Frequenz Gold-standard (n=10575)	Absolute Fre-quenz RIC (n=9378)	Prüf-größe	p	KI (95%)
Abkürzungen, Maßeinheiten						
др$_F$	a. (andere)	7	0	-	0,0168	[1,2786, ∞)[97]
мм	mm	15	0	11,4909	0,0007	[0,0006, 0,0022]
рис	Abb. (Zeichnung)	14	0	10,6076	0,0011	[0,0005, 0,0021]
т	-	21	0	16,8016	< 0,0001	[0,0010, 0,0029]
Adjektive/Adverbien						
наиболее$_F$	Adverb für die Superlativ-bildung	9	1	-	0,0237	[1,1064, 349,5879]
обычно	normaler-weise, gewöhnlich	11	0	7,9637	0,0048	[0,0003, 0,0018]
Nomina						
вид*	Art, Spielart, Typ; Form	25	4	11,5562 (6,4663)	0,0007 (0,0110)	[0,0008, 0,0031]
деталь	Bauteil; Werkstück	21	0	16,8016	< 0,0001	[0,0010, 0,0029]
зависимость	Abhängigkeit	18	1	11,6754	0,0006	[0,0007, 0,0025]
изготовление$_F$	Herstellung	7	0	-	0,0168	[1,2786, ∞)
изделие$_F$	Produkt; Bauteil; Werkstück	9	0	-	0,0044	[1,7513, ∞)
использование	Gebrauch, Benutzung	14	1	8,2501	0,0041	[0,0004, 0,0020]
конструкция	Konstruktion	19	0	15,0297	0,0001	[0,0009, 0,0027]
материал	Material	14	2	6,3281	0,0119	[0,0003, 0,0020]
поверхность	Oberfläche	22	0	17,6881	< 0,0001	[0,0011, 0,0030]
помощь	Hilfe	18	1	11,6754	0,0006	[0,0007, 0,0025]
свойство	Eigenschaft	14	0	10,6076	0,0011	[0,0005, 0,0021]
совокупность*$_F$	Gesamtheit	7	0	-	0,0168 (0,0179)	[1,2786, ∞)

[97] Diese Notation bezeichnet ein sogenanntes halboffenes, d. h. im hier dargestellten Fall ein nur nach unten beschränktes Intervall.

способ	Methode, Mittel	14	3	4,7649	0,0290	[0,0001, 0,0019]
тип	Typ	13	1	7,4053	0,0065	[0,0003, 0,0019]
Präpositionen						
для	für	139	44	38,1504	< 0,0001	[0,0058, 0,0111]
из*	aus	104	40	20,7457 (6,6737)	< 0,0001 (0,0098)	[0,0032, 0,0080]
от	von	65	35	5,3363	0,0209	[0,0004, 0,0044]
под	unter	26	10	4,6047	0,0319	[0,0001, 0,0026]
при	bei	90	10	53,7532	< 0,0001	[0,0055, 0,0094]
Verben						
использовать	benutzen	30	4	15,5877	0,0001	[0,0012, 0,0036]
называть	nennen	15	3	5,4919	0,0191	[0,0002, 0,0020]
называться*$_F$	heißen	8	0	-	0,0086 (0,0450)	[1,5144, ∞)
обеспечивать*	gewährleisten	27	2	17,1736 (5,4834)	< 0,0001 (0,0192)	[0,0012, 0,0034]
образовывать$_F$	bilden	7	0	-	0,0168	[1,2786, ∞)
поддерживать$_F$	unterstützen; sorgen für, aufrecht erhalten	6	0	-	0,0329	[1,0444, ∞)
подразделяться*$_F$	unterteilt werden	9	0	-	0,0044 (0,0258)	[1,7513, ∞)
предназначить*	vorsehen, bestimmen	26	0	21,2365 (13,3241)	< 0,0001 (0,0003)	[0,0014, 0,0035]
представлять*	darstellen	31	2	20,6251 (17,3121)	< 0,0001 (< 0,0001)	[0,0015, 0,0039]
применять*	anwenden, verwenden	20	2	11,2286 (4,5444)	0,0008 (0,0330)	[0,0007, 0,0027]
различать*$_F$	unterscheiden	8	0	-	0,0086 (0,0450)	[1,5144, ∞)
служить*	dienen	27	1	19,5202 (12,2535)	< 0,0001 (0,0005)	[0,0014, 0,0035]
состоять*	bestehen (aus)	57	2	43,4414 (26,3396)	< 0,0001 (< 0,0001)	[0,0036, 0,0067]
устанавливать*	aufstellen; festlegen; installieren	36	0	30,1205 (8,6494)	< 0,0001 (0,0033)	[0,0022, 0,0046]

устанавливаться$_F$	aufgestellt; installiert; festgelegt werden	8	0	-	0,0086	[1,5144, ∞)

Tabelle 42: Kandidaten für russische Boostwörter

Die Untersuchung zeigt, dass für eine Vielzahl derjenigen Boostwort-Kandidaten, die sich aus dem Vergleich zwischen den KRC-Gold-Daten und der RIC-Stichprobe ergeben, für den Vergleich mit den Extraktionskorpora Sub-EKR und Sub-TKR die angenommenen Verteilungsunterschiede nicht bestätigt werden können. Mithin scheint es sich hierbei um Lemmata zu handeln, die für eher technisch-fachliche Texte als spezifisch eingeschätzt werden können, eine diskriminatorische Kraft hinsichtlich der Abgrenzung wissenshaltiger Kontexte von generischen Sätzen der Extraktionskorpora kann ihnen aber nicht unterstellt werden. Für diese Annahme spricht auch, dass es sich bei den betreffenden Einheiten um technische Abkürzungen, Nominalisierungen und relativ abstrakte Nomina handelt, die verschiedene Aspekte von Gerätebauteilen benennen (z. B. *изготовление*, „Herstellung“, *материал,* “Material“, *поверхность,* „Oberfläche“, *свойство,* „Eigenschaft“, *тип*, „Typ“) oder entsprechende Phrasen einleiten (*помощь -> с помощью,* „mit Hilfe von“), die in den Gold-Daten zwar häufiger erscheinen als in der RIC-Stichprobe, jedoch auch in der Stichprobe aus den Extraktionskorpora mit ähnlichen Anteilen vertreten sind. Als generische Nomina bestätigt der Vergleich der Gold-Daten mit den Extraktionskorpora lediglich die Nomina *вид* und *совокупность.* Das Nomen *вид* („Art, Typ, Spielart“) wird dabei verwendet, um auf Begriffsklassen zu verweisen und scheint insbesondere auch für die Darstellung von Hyponymie/Hyperonymie-Beziehungen eine Rolle zu spielen. Allerdings erscheint das Nomen auch als Bestandteil der Phrase *в виде* („in Form von, in Gestalt von“) und dient dann zur Beschreibung der Form oder des Aufbaus von Gegenständen oder Anlagen, vgl. Beispiele 18) und 19):

18) Контактная сварка относится к <u>видам</u> сварки с кратковременным нагревом места соединения без оплавления или с оплавлением и осадкой разогретых заготовок.
[*Übersetzung*: Das Widerstandschweißen gehört <u>zu jenen Schweißverfahren</u>, bei denen die Verbindungsstelle ohne Abbrennen kurz erhitzt wird oder die erhitzten Schweißteile beim Abbrennen gepresst werden.]

19) Зубча́тое колесо́ (шестерня́) — основная деталь зубчатой передачи <u>в виде</u> диска с зубьями на цилиндрической или конической поверхности, входящими в зацепление с зубьями другого зубчатого колеса.
[*Übersetzung:* Das Zahnrad ist ein wesentlicher Bestandteil eines Zahnrad-Getriebes <u>in Gestalt</u> einer Scheibe, auf deren zylindrischer oder konischer Oberfläche Zähne angebracht sind, welche in die Zähne eines anderen Zahnrads greifen.]

Das Nomen *совокупность* („Gesamtheit, Einheit“) scheint dagegen häufig als *genus proximum* zu fungieren, vgl. Beispiele 20) und 21).

20) Конструктивно АБС представляет собой <u>совокупность</u> датчиков, модуляторов и блока управления.
[*Übersetzung*: Aus konstruktiver Sicht stellt das ABS eine <u>Baugruppe</u> dar, die aus Sensoren, Modulatoren und dem Steuerblock besteht.]

21) Передаточный механизм – совокупность устройств, которая в соответствии с командами органа управления передает энергию от источника или аккумулятора исполнительным органам привода.
[*Übersetzung*: Als Getriebe wird diejenige Baugruppe bezeichnet, die in Abhängigkeit von den Anweisungen der Steuereinheit die Energie von der Energiequelle oder dem Energiespeicher an die Ausführungseinheiten des Antriebs überträgt.]

Das gehäufte Auftreten der Präposition *из* („aus“, *состоит из, изготовлен из*) kann möglicherweise auf die ebenfalls gesteigerte Häufigkeit des Verbs *состоять* („bestehen aus“) im Goldstandard, das diese Präposition verlangt, zurückgeführt werden.
Der Vergleich der Gold-Daten mit dem Extraktionskorpus erbringt auch eine Reihe neuer Boostwort-Kandidaten, und zwar das Relativpronomen *который* (Prüfgröße 7,0817, $p = 0{,}0078$), das Adverb *непосредственно* („unmittelbar“, $p = 0{,}0345$), die Konjunktion *и* („und“, Prüfgröße 4,9900, $p = 0{,}0255$) und das Reflexivpronomen *себя* (Prüfgröße 21,6945, $p < 0{,}0001$). Bereits Pearson (1998: 141, vgl. auch Abschnitt 3.2) hatte darauf verwiesen, dass Angaben von Begriffseigenschaften häufig durch Relativsätze eingeleitet werden. Diese Tendenz scheint sich in Ansätzen auch für das Russische zu bestätigen (vgl. Beispiele 22) und 23)), allerdings findet sich das Pronomen auch in weniger restriktiven Zusammenhängen.

22) Аккумулятор энергии – устройство, которое накапливает энергию, предназначенную для осуществления торможения.
[*Übersetzung*: Ein Akkumulator ist ein Gerät, das die für den Bremsvorgang vorgesehene Energie speichert.]

23) В настоящее время получает распространение воздушная система охлаждения, при которой необходимый температурный режим двигателя создается воздушным потоком.
[*Übersetzung*: Immer weitere Verbreitung finden gegenwärtig Luftkühlsysteme, in denen die Regelung der Motortemperatur durch einen Luftstrom erfolgt.]

Das Adverb *непосредственно* taucht in unterschiedlichsten Kontexten auf, aufgrund seiner insgesamt niedrigen Häufigkeit lassen sich jedoch keine gesicherten Schlüsse ziehen. Das gehäufte Auftreten der Konjunktion *и* im Goldstandard scheint dagegen auf das bereits für das Deutsche festgestellte Überwiegen nebenordnender Strukturen hinzudeuten. Das Reflexivpronomen *себя* dagegen tritt höchstwahrscheinlich insbesondere gemeinsam mit dem Verb *представлять* (*представлять собой*, „darstellen“) auf und bildet mit diesem ein den in der einschlägigen Literatur beschriebenen ähnelndes *knowledge pattern*.
Die Mehrzahl der in Tabelle 42 präsentierten Verben lässt sich insofern aus der in Tabelle 4 (Abschnitt 3.2.3) dargestellten Zusammenfassung lexikalisch-syntaktischer Muster in existierenden Studien zu anderen Sprachen ableiten, als sie sowohl den dort angeführten semantischen Relationen als auch den semantischen Paradigmen der anderssprachigen Muster zugeordnet werden können: Sie stellen mithin Musterkandidaten für die Extraktion russischer wissenshaltiger Kontexte dar. Von Interesse ist zudem, dass die Verben in Tabelle 42 mit Ausnahme von *предназначить* imperfektiv sind, was als Generalisierungssignal gewertet werden kann. Der imperfektive Aspekt wird im Russischen entweder zur Bezeichnung von Prozessen (*имперфективно-процессуальная функция,*

„imperfektiv-prozessuale Funktion“) oder aber gewohnheitsmäßig wiederholten Vorgängen (*узуально-характеризующая функция,* „usuell-charakterisierende Funktion“, Zolotova et al. 1998: 27, Hervorhebungen durch die Autorin) verwendet:

> „В имперфективно-процессуальной функции действие или состояние предстают в их наблюдаемой протяженности, не ограниченной временными рамками [...]. В имперфективной узуально-характеризующей функции говорящий с дистанцированной во времени и пространстве позиции представляет действия, состояния и отношения как обычные занятия, умения, характеристики [...].“
> [*Übersetzung:* „In der imperfektiv-prozessualen Funktion werden eine Handlung oder ein Zustand in ihrer beobachtbaren, nicht durch temporale Schranken begrenzten Ausdehnung betrachtet […]. In der imperfektiven usuell-charakterisierenden Funktion stellt der Sprecher unter einem sowohl zeitlich als auch räumlich distanzierten Blickwinkel Handlungen, Zustände und Beziehungen als übliche Beschäftigungen, Fähigkeiten und Eigenschaften dar […].“]

Konjugierte Formen perfektiver russischer Verben beschreiben dagegen eine in der Zukunft liegende Handlung im Hinblick auf ihre temporale Abgeschlossenheit (*перфективная функция*, „perfektive Funktion“) bzw. als Folge von Einzelhandlungen (*аористная функция*, „Aorist-Funktion“, ebd.)[98] und kommen entsprechend in wissenshaltigen Kontexten, die Ist-Zustände in allgemein gültiger Form beschreiben, nur selten vor. Das perfektive Verb *предназначить* erscheint nun im Goldstandard überwiegend als prädikativ gebrauchtes Adjektiv in Kurzform, eine Verwendung, die grammatisch dem deutschen Zustandspassiv (vgl. Zifonun et al. 1997: 1791) ähnelt und damit ebenfalls einen zum Zeitpunkt der Äußerung gültigen Zustand beschreibt, vgl. Beispiele 24) und 25):

24) Воздушный фильтр предназначен для очистки воздуха, поступающего от компрессора в пневматическую систему от влаги и от масла.
[*Übersetzung:* Der Luftfilter dient zur Reinigung der vom Kompressor in die Pneumatik gelangenden Luft von Feuchtigkeit und Öl.]

25) Главный тормозной цилиндр предназначен для преобразования усилия, прилагаемого к педали тормоза, в избыточное давление тормозной жидкости и распределения его по рабочим контурам.
[*Übersetzung*: Der Hauptbremszylinder dient zur Umwandlung der auf das Bremspedal einwirkenden Kraft in einen Überdruck der Bremsflüssigkeit und dessen Verteilung auf die Arbeitskreise.]

Die Verben *называться, подразделяться* und *устанавливаться* sind Passivformen regulärer Verben und stützen in Übereinstimmung mit den bisherigen Ergebnissen für das Deutsche die Annahme eines häufigen Gebrauchs von Passivformen in wissenshaltigen Kontexten[99]. Eine eingehendere Diskussion von lexikalischen Bausteinen für *knowledge patterns* findet sich in Abschnitt 5.2.2.5. Die in Tabelle 42 aufgeführten Verben sollen daher an dieser Stelle nicht näher diskutiert werden.

[98] Für eine sehr differenzierte und auf den deutschen Russisch-Lerner zugeschnittene Darstellung des russischen Verbalaspekts sei auf Bendixen/Voigt (1999: 85 ff.) verwiesen.

[99] Genauere Analysen im Hinblick auf Aspektzuordnungen und Passivgebrauch folgen in späteren Abschnitten.

In Bezug auf die im Merkmalskatalog abgefragten lexikalischen Merkmale kann nunmehr festgestellt werden, dass die hier beschriebene Analyse sie für das Russische überwiegend nicht bestätigen konnte. Weder Frequenzadverbien noch generische Nomina wie die in Literaturbeispielen für andere Sprachen genannten „Begriff“ (*понятие*) oder „Terminus“ (*термин*) konnten für den russischen Goldstandard als Boostwörter identifiziert werden. Auch ein russisches Analogon zu den von Cramer (2011) genannten definitionstypischen Adverbien konnte nicht gefunden werden, wobei einzuräumen ist, dass hierzu möglicherweise eine Untersuchung auf phraseologischer Ebene erforderlich wäre[100]. Ein russisches Analogon zu den unpersönlichen Personalpronomen „man“ bzw. frz. *on* existiert nicht. Vielmehr werden Verallgemeinerungen dieser Art im Russischen durch einen sogenannten „unbestimmt-persönlichen“, subjektlosen Satz ausgedrückt, in dem das Prädikat die dritte Person Plural annimmt (vgl. Bendixen/Voigt 1999: 156, Kirschbaum 2001: 65)[101]. Die identifizierten generischen Nomina unterscheiden sich ebenfalls von den in der Literatur angeführten. So beziehen sie sich nicht auf Bestandteile der Definition bzw. den terminologischen Status‘ des Definiendums („Begriff“), sondern fungieren als Oberbegriffe oder verweisen auf wesentliche Aspekte des Definieren und Erklärens, nämlich die Abgrenzung von Analyseebenen, die Erläuterung von Hierarchien und fachlichen Zusammenhängen.

Grundsätzlich ist darauf hinzuweisen, dass die bisherige Untersuchung nur Tendenzen beschreiben kann – die geringen Häufigkeiten der identifizierten „Boostwörter“ machen deutlich, dass für stabile Ergebnisse wesentlich größere Datenmengen als die hier zur Verfügung stehenden benötigt werden. Auch die im Zusammenhang mit dem Fisher-Test ermittelten halboffenen Konfidenzintervalle zeigen, dass die tatsächlichen Verteilungen der „Boostwörter“ in zwei angenommenen größeren Grundgesamtheiten aufgrund der gegebenen Datenmengen nicht zuverlässig ermittelt werden können. Im Hinblick auf das in dieser Arbeit verfolgte Extraktionsvorhaben ist zudem angesichts der niedrigen Frequenzen der Lemmata und der geringen geschätzten Größen der Verteilungsunterschiede fraglich, inwieweit die identifizierten „Boostwörter“ für praktische Extraktionsaufgaben tatsächlich relevant sind, d. h. inwieweit sie zur Verbesserung von Extraktionsergebnissen beitragen können. Untersuchungen auf Grundlage von größeren Datenmengen könnten auch dahingehend von Interesse sein, als mögliche Lemmatisierungsfehler durch die Fülle des analysierbaren Materials ausgeglichen und mithin mehr „Boostwörter“ ermittelt werden könnten. Tabelle 43 fasst analog zu den bisherigen Ausführungen die Ergebnisse für mögliche Stopwörter zusammen.

Wort	**Übersetzung**	**Absolute Frequenz Gold-standard (n=10575)**	**Absolute Frequenz RIC (n=9378)**	**Prüf-größe**	**p**	**KI (95%)**
Adjektive						
весь	der ganze	18	40	10,3994	0,0013	[-0,0042, -0,0009]
все	alle	0	15	14,8650	0,0001	[-0,0025, -0,0007]
любой	jeder beliebige	4	14	5,6701	0,0173	[-0,0021, -0,0001]

[100] Phraseologische Aspekte werden in Abschnitt 5.2.3.6 diskutiert.
[101] In Bezug auf den Goldstandard wird dieser Aspekt in Abschnitt 5.2.3.2.1 erörtert.

настоящий$_F$	echt; vorliegend	1	8	-	0,0158	[0,0025, 0,8265]
новый*	neu	1	14	11.1422 (-)	0,0008 (0,0144)	[-0,0023, -0,0005]
последний$_F$	der letzte, Letzterer	1	7	-	0,0301	[0,0028, 9858]
чрезвычай-ный$_F$	außerordent-lich	0	5	-	0,0229	[0, 0,9675]
Adverbien						
возможно	möglich	0	5	-	0,0229	[0, 0,9675]
вообще$_F$	allgemein, überhaupt	0	5	-	0,0229	[0, 0,9675]
вот$_F$	hier, da (Deixis)	0	8	-	0,0024	[0, 0,5193]
даже*	sogar	0	13	12,6171 (-)	0,0004 (0,0009)	[-0,0022, -0,0005]
здесь$_F$	hier	0	6	-	0,0108	[0, 0,7529]
еще	noch, schon	3	12	5,3036	0,0213	[-0,0019, -0,0001]
лишь$_F$	lediglich, nur	0	6	-	0,0108	[0, 0,7529]
очень*$_F$	sehr	1	8	-	0,0158 (0,0144)	[0,0025, 0,8265]
пока$_F$	solange; tschüß	0	6	-	0,0108	[0, 0,7529]
сам	selbst	0	12	11,4946	0,0007	[-0,0021, -0,0005]
сегодня	heute	1	10	6,8460	0,0089	[-0,0018, -0,0002]
там$_F$	dort	0	7	-	0,0051	[0, 0,6150]
только	nur	5	21	10,5991	0,0011	[-0,0029, -0,0006]
уже$_F$	schon	0	10	-	0,0005	[0, 0,3953]
хорошо	gut	0	5	-	0,0229	[0, 0,9675]
Konjunktionen						
а	aber	36	66	12,1977	0,0005	[-0,0058, -0,0015]
если*	wenn, falls	3	27	20,6059 (19,6255)	< 0,0001 (< 0,0001)	[-0,0038, -0,0014]
но	aber, jedoch	5	35	24,7872	< 0,0001	[-0,0047, -0,0019]
однако	jedoch, allerdings	2	14	8,9794	0,0027	[-0,0022, -0,0004]
чтобы	damit, um zu	7	17	4,5631	0,0327	[-0,0022, -0,0001]

Nomen						
вопрос	Frage	1	11	7,9062	0,0049	[-0,0019, -0,0003]
мнение$_F$	Meinung	0	5	-	0,0029	[0, 0,9675]
необходи-мость*	Notwendig-keit	0	6	-	0,0108 (0,0042)	[0, 0,7529]
Partikel						
бы	Irrealis-partikel	1	20	17,7463	< 0,0001	[-0,0031, -0,0010]
ли$_F$	Fragepartikel	0	10	-	0,0005	[0, 0,3953]
не*	nicht	20	112	74,8915 (18,3885)	< 0,0001 (< 0,0001)	[-0,0125, -0,0076]
нет*	nein	0	17	17,1159 (-)	< 0,0001 (0,0202)	[-0,0028, -0,0009]
ни$_F$	weder, nicht-	0	5	-	0,0229	[0, 0,9675]
Präpositionen						
в	in	333	345	4,0916	0,0431	[-0,0105, -0,000]
о	über	6	39	26,9143	< 0,0001	[-0,0051, -0,0021]
по	~ relativ zu	63	84	5,7123	0,0168	[-0,0055, -0,0005]
после*	nach	0	21	21,6233 (10,9239)	< 0,0001 (0,0009)	[-0,0033, -0,0012]
Pronomen						
ваш$_F$	euer, Ihr	0	8	-	0,0024	[0, 0,5193]
вы*	ihr, Sie	0	25	26,1362 (-)	< 0,0001 (0,0202)	[-0,0038, -0,0015]
ее*	sie (fem. Gen./Akk.)	4	14	5,6701 (4,3801)	0,0173 (0,0364)	[-0,0021, -0,0001]
как*	wie	25	63	20,4765 (5,8738)	< 0,0001 (0,0154)	[-0,0063, -0,0024]
кто$_F$	wer	0	10	-	0,0005	[0, 0,3953]
мы	wir	3	50	45,9213	< 0,0001	[-0,0067, -0,0034]
наш	unser	2	23	18,5797	< 0,0001	[-0,0034, -0,0011]
он	er	7	25	11,2446	0,0008	[-0,0033, -0,0008]
она	sie	8	21	6,5427	0,0105	[-0,0027, -0,0003]
они	sie (Pl.)	13	39	15,3007	0,0001	[-0,0045, -0,0014]
свой	sein, das eigene	5	33	22,6852	< 0,0001	[-0,0044, -0,0017]

так*	so	14	48	21.8936 (3,9389)	< 0,0001 (0,0472)	[-0,0055, -0,0021]
такой	so einer, ein solcher	11	27	7,9010	0,0049	[-0,0032, -0,0005]
тот	jener	10	30	11,5130	0,0007	[-0,0036, -0,0009]
ты	du	0	12	11,4946	0,0007	[-0,0021, -0,0005]
это*	dieses; das ist	4	53	46,6862 (13,2017)	< 0,0001 (0,0003)	[-0,0069, -0,0036]
этот*	dieser	24	59	18,4487 (7,7661)	< 0,0001 (0,0053)	[-0,0060, -0,0021]
что*	was	12	100	79,1494 (21,0098)	< 0,0001 (< 0,0001)	[-0,0118, -0,0073]
я	ich	0	74	81,634	< 0,0001	[-0,0098, -0,0060]
Verben						
быть*	sein	23	88	45,3942 (16,6315)	< 0,0001 (< 0,0001)	[-0,0087, -0,0045]
говорить	sprechen	0	12	11,4946	0,0007	[-0,0021, -0,0005]
делать	tun, machen	0	12	11,4946	0,0007	[-0,0021, -0,0005]
должен*	er muss	0	26	27,2651 (16,2673)	< 0,0001 (0,0001) [102]	[-0,0039, -0,0016]
заявить$_F$	erklären, kundgeben	0	5	-	0,0229	[0, 0,9675]
знать	wissen	0	11	10,3734	0,0013	[-0,0020, -0,0004]
идти$_F$	gehen	0	8	-	0,0024	[0, 0,5193]
мочь*	können	15	32	7,5807 (3,8459)	0,0059 (0,0499)	[-0,0035, -0,0005]
пользоваться	benutzen	0	11	10,3734	0,0013	[-0,0020, -0,0004]
понять$_F$	verstehen	0	5	-	0,0229	[0, 0,9675]
сделать	tun, machen	0	11	10,3734	0,0013	[-0,0020, -0,0004]
сказать$_F$	sagen	0	7	-	0,0051	[0, 0,6150]
создать$_F$	schöpfen, schaffen, erstellen	1	7	-	0,0301	[0,0028, 0,9858]
становиться$_F$	werden	0	5	-	0,0229	[0, 0,9675]

[102] Abweichend vom vorherigen Experiment wird für den Vergleich *Goldstandard-Extraktionskorpora* die lemmatisierte Form *должный* zugrunde gelegt.

считать$_F$	denken; halten für; zählen	0	8	-	0,0024	[0, 0,5193]
хотеть	wollen	0	11	10,3734	0,0013	[-0,0020, -0,0004]

Tabelle 43: Kandidaten für russische Stopwörter

Zusätzlich zu den in der Tabelle zusammenfassend dargestellten Stopwörtern ergibt der Vergleich der KRC-Gold-Daten mit den Extraktionskorpora Sub-EKR und Sub-TKR Verteilungsunterschiede für weitere Adverbien (*необходимо,* „erforderlich“, p = 0,0009, *почти,* „fast“, p = 0,0092, *мало*, „wenig“, p = 0,0530), Adjektive (*многий,* „viel“, Prüfgröße 7,2890, p = 0,0069, *сложный,* „kompliziert“, p = 0,0092), Verben (*установить,* „aufstellen“, Prüfgröße 5,3797, p = 0,0204, *следовать*, „folgen; sollen“, Prüfgröße 4,3870, p = 0,0362), Pronomen (*чем,* hier wahrscheinlich in der Bedeutung einer Konjunktion: „als“ in Vergleichen, Prüfgröße 4,8065, p = 0,0284) und Konjunktionen (*кроме,* „außer“, p = 0,0144, *поэтому*, „deshalb“, p = 0,0514). Im Ergebnis dieses zweifachen Vergleichs zeigt sich, dass die Goldstandard-KRCs zwar einerseits vom RIC durch solche Stopwörter abgegrenzt werden kann, die eher der Allgemeinsprache zuzuordnen sind. Andererseits weisen beide Vergleiche auf Merkmale hin, die für die KRCs als eher untypisch angesehen werden können, d. h. durch deren Abwesenheit oder schwache Ausprägung sie generell gekennzeichnet zu sein scheinen. Hier sind einerseits modale Ausdrücke – Modalverben und Prädikative wie *должен, мочь, необходимо,* und *следовать* – zu nennen, die der Goldstandard kaum verwendet. Ähnlich, nämlich als Hinweis auf einen faktenorientierten Stil mit nur wenigen offen subjektiven Elementen kann auch der weitgehende Verzicht auf verstärkende Adverbien wie *очень* und *даже* sowie qualifizierende Adjektive interpretiert werden. Auf eine Tendenz zur Formulierung expliziter Aussagen deutet der mit Cramers Beschreibung alltagssprachlicher deutscher Definitionen übereinstimmende weitgehende Verzicht der Goldstandard-KRCs auf Pronomen unterschiedlichster Art zu, wobei der Verzicht auf Demonstrativa ebenso wie das seltene Auftreten von Konjunktionen mit konditionaler Bedeutung (*если*) auf ein Streben nach Allgemeinheit, d. h. die Loslösung von einzelfallabhängigen Bedingungen hindeutet. Das seltene Auftreten des Verbs *быть* („sein“) dagegen ist ein Indiz für die überwiegende Verwendung des Tempus Präsens, da „sein“ in russischen Präsenssätzen nicht ausgedrückt (dies wird als Nullkopula bezeichnet, alternativ kann auch ein Gedankenstrich gesetzt werden, vgl. Beispiele 19), 21) und 22)) und nur zur analytischen Bildung des Futurs (imperfektiver Aspekt) sowie im Präteritum verwendet wird[103]. Verzichtet und damit einem expliziten Selektionskriterium für wissenshaltige Kontexte entsprochen wird desgleichen auf Elemente mit anaphorischer Bedeutung, d. h. lexikalische Einheiten, die Rückbezüge auf bereits eingeführtes Wissen herstellen und damit auf die Diskursebene verweisen. Als Beispiele zu nennen sind hier etwa die Einheiten *поэтому,* und *кроме* (*того,* „außerdem“). Nicht zuletzt entspricht auch der Verzicht auf Verneinung linguistischen Intuitionen über wissenshaltige Kontexte. Tendenziell scheint die Untersuchung damit die Annahmen Walters (2010, vgl. Abschnitt 3.2) zu deutschen Definitionen auch für russische wissenshaltige Kontexte zu stützen, allerdings besteht auch hier Grund zu der Annahme, dass gesicherte Aussagen die Untersuchung größerer Datenmengen voraussetzen. Zudem ist auch in Bezug auf die Analyse der Stopwörter wieder auf die bereits beschriebene Lemmatisierungsproblematik zu verweisen.

[103] Auch in Bezug auf dieses Thema sei auf den Abschnitt zur Verbmorphologie verwiesen.

5.2.2.3.2 Deutsch

Analog zur Darstellung für das Russische sollen nun auch Stop- und Boostwörter für deutsche KRCs präsentiert werden. Tabelle 44 fasst dabei zunächst die Boostwörter zusammen, die nach Überprüfung mit Hilfe des Chi-Quadrat-Tests in den KRCs unterschiedliche Verteilungen im Vergleich zum NCL aufweisen, übereinstimmende Ergebnisse für den Vergleich mit der Stichprobe aus den Extraktionskorpora Sub-EKD und Sub-TKD werden wiederum mit * markiert. Auch hier wird in der Diskussion der Beispiele im Sinne eines leichteren Verständnisses die interessierende lexikalische Einheit unterstrichen.

Wort	**Absolute Frequenz Gold-standard (n=12073)**	**Absolute Frequenz NCL (n=9265)**	**Prüf-größe**	**p**	**KI (95%)**
Abkürzungen					
bzw$_F$	11	0	-	0,0035	[1,9272, ∞)
Adjektive					
modern*	15	2	5,7100 (6,1485)	0,0169 (0,0132)	[0,0002, 0,0018]
sogenannt*	32	1	20,3314 (18,8559)	< 0,0001 (< 0,0001)	[0,0015, 0,0036]
Adverbien					
direkt	22	2	10,6529	0,0011	[0,0007, 0,0025]
ebenfalls$_F$	10	1	-	0,0292	[1,0921, 332,8408]
gering$_F$	9	0	-	0,0065	[1,5154, ∞)
häufig$_F$	9	0	-	0,0065	[1,5154, ∞)
hoch	27	4	10,5574	0,0012	[0,0008, 0,0028]
Artikel					
die*	1510	972	20,5348 (7,3669)	< 0,0001 (0,0066)	[0,0115, 0,0288]
eine*	417	206	27,5736 (10,5794)	< 0,0001 (0,0011)	[0,0078, 0,0168]
Konjunktionen					
entweder*$_F$	8	0	-	0,0118 (0,0080)	[1,3104, ∞)
oder*	107	18	41,9231 (23,3219)	< 0,0001 (< 0,0001)	[0,0049, 0,0089]
wobei$_F$	7	0	-	0,0215	[1,1064, ∞)
Nomen					
Begriff*$_F$	10	1	-	0,0292 (0,0020)	[1,0921, 332,8408]
Form*	20	0	13,6446 (10,2979)	0,0002 (0,0013)	[0,0008, 0,0025]
Partikel					
dar	14	1	6,8244	0,0090	[0,0003, 0,0018]
Präpositionen					
als	87	40	6,9141	0,0086	[0,0008, 0,0050]

aus*	116	34	25,6402 (26,9639)	< 0,0001 (< 0,0001)	[0,0037, 0,0082]
bei	115	37	21,9057	< 0,0001	[0,0033, 0,0078]
durch	88	15	33,9121	< 0,0001	[0,0039, 0,0075]
mit	150	78	7,5823	0,0059	[0,0012, 0,0068]
von	159	84	7,4802	0,0062	[0,0012, 0,0070]
zwischen*	33	3	16,6683 (8,9403)	< 0,0001 (0,0028)	[0,0013, 0,0035]
Pronomen					
welche*	22	4	7,2251 (6,0331)	0,0072 (0,0140)	[0,0004, 0,0024]
Verben					
befinden*	17	2	7,0891 (9,5858)	0,0078 (0,0020)	[0,0004, 0,0020]
beschreiben	12	0	7,5306	0,0061	[0,0003, 0,0017]
besitzen	12	0	7,5306	0,0061	[0,0003, 0,0017]
bestehen*	54	3	32,3319 (38,9082)	< 0,0001 (< 0,0001)	[0,0028, 0,0055]
$bestimmt_F$	7	0	-	0,0215	[1,1064, ∞)
bezeichnen*	40	0	29,0090 (30,6983)	< 0,0001 (< 0,0001)	[0,0022, 0,0044]
dienen*	14	0	9,0550 (5,3607)	0,0026 (0,0206)	[0,0006, 0,0022]
einsetzen*	22	1	12,7602 (7,2574)	0,0004 (0,0071)	[0,0008, 0,0026]
erzeugen	13	0	8,2923	0,0040	[0,0004, 0,0018]
genannt*	17	1	9,0276 (11,7179)	0,0027 (0,0006)	[0,0005, 0,0021]
nennen*	28	2	15,0546 (11,8834)	0,0001 (0,0006)	[0,0011, 0,0031]
nutzen*	25	2	12,8416 (9,5184)	0,0003 (0,0020)	[0,0009, 0,0028]
unterscheiden*	18	1	9,7694 (10,4684)	0,0018 (0,0012)	[0,0006, 0,0022]
verwenden	18	2	7,7906	0,0053	[0,0004, 0,0021]
werden*	257	105	30,5515 (6,1049)	< 0,0001 (0,0135)	[0,0065, 0,0134]

Tabelle 44: Kandidaten für deutsche Boostwörter

Zusätzlich zu den in der Tabelle zusammengefassten Daten ergab der Vergleich des KRC-Goldstandards mit der Stichprobe aus den Extraktionskorpora signifikante Verteilungsunterschiede für die Lemmata „kein“ (Prüfgröße 11,3924, p = 0,0007, wobei allerdings Lemmatisierungsfehler nicht ausgeschlossen werden können: im Vergleich zum NCL erscheinen „kein“ und „keine“ deutlich seltener im Goldstandard und „keine“ ist auch im Hinblick auf die Extraktionskopora ein Stopwort), „all“ (p = 0,0020), „mehrer“ (p = 0,0040), „jed“ (p = 0,0159), „angeordnet“ (p = 0,0316, hier ergibt auch der Vergleich zum NCL ein signifikantes Ergebnis, jedoch kann aufgrund der sehr niedrigen Frequenz des Lemmas nicht eindeutig abgeschätzt werden, welche Richtung der Verteilungsunterschied hat, das Konfidenzintervall schwankt entsprechend zwischen einem Wert unter 1 und ∞, so dass das Lemma in der Tabelle nicht berücksichtigt wurde) und „gemeinsam“ (p =

0,0396). Im Hinblick auf die in der Tabelle angeführten Daten ist ebenfalls auf die geringen Frequenzen und geringen Größen der Verteilungsunterschiede hinzuweisen, was bedeutet, dass die dargestellten lexikalischen Merkmale als eher marginal einzustufen sind. Die nicht durch den Vergleich mit dem Extraktionskorpus gestützten Verteilungsunterschiede können dabei wiederum als eher typisch für technische Texte interpretiert werden, scheinen jedoch nicht oder kaum zur Unterscheidung wissenshaltiger Kontexte von generischen Sätzen beizutragen. Im Gegensatz dazu scheinen die allein im Vergleich mit den Extraktionskorpora ermittelten Kandidaten für Boostwörter „all“ und „mehrer“ auf generalisierende Aussagen hinzudeuten, womit eine wissenshaltigen Kontexten unterstellte Eigenschaft realisiert wäre. Das Adjektiv „modern“ scheint dazu zu dienen, eine Aussage hinsichtlich ihrer zeitlichen Gültigkeit zu kennzeichnen, es erfüllt mithin eine aktualisierende Funktion, vgl. Beispiele 26) und 27):

26) Moderne Kohlekraftwerke verfügen daher über Vorrichtungen zur Entschwefelung, Entstickung und Entstaubung sowie zum Rückhalt der toxischen Spurenstoffe.

27) Moderne Vertikal-Achser bezeichnet man unter anderem auch als H-Rotor, weil die Blätter senkrecht zur Achse oder Rotorwelle stehen.

Das Adjektiv „sogenannt“ erschien bereits in Tabelle 4 als deutsches Extraktionsmuster. In den untersuchten wissenshaltigen Kontexten dient es zur Fokussierung auf ein neues Thema, in der Regel das Definiendum.

28) Das sogenannte Aufwindkraftwerk basiert auf dem Grundgedanken, eine Luftströmung wie in der Natur durch Erwärmung, das heißt Unterschiede in der Luftdichte, herbeizuführen.

29) Der sogenannte Abstandsfaktor kA beschreibt den zwischen zwei benachbarten Konvertern minimal einzuhaltenden Abstand.

Das deutlich häufigere Auftreten von Artikeln im KRC-Goldstandard stützt die bisherigen Ausführungen zur größeren Häufigkeit von Nominalphrasen, wobei der Umstand, dass sowohl unbestimmte als auch bestimmte Artikel als Boostwörter im Sinne der hier entwickelten Definition identifiziert wurden, darauf hindeutet, dass eine deutliche Systematik des Artikelgebrauchs in deutschen wissenshaltigen Kontexten nicht vorliegt. Das deutlich häufigere Auftreten der Konjunktion „oder“ stützt wiederum die Annahme der überwiegenden Verwendung parataktischer Strukturen. Die Konjunktion „entweder“ wird dagegen zur Spezifikation konkurrierender Varianten verwendet und kann entsprechend auch komplementäre Begriffe zueinander in Beziehung setzen:

30) Von der Stellung der Windräder kann man solche unterscheiden, die entweder in Luv (der Wind bläst von vorne auf die Gondel, bzw. die Rotorblätter) oder in Lee (der Wind bläst von hinten auf die Gondel bzw. die Rotorblätter) stehen.[104]

Das Nomen „Begriff“ bestätigt dagegen eine bereits dargestellte Beobachtung Walters (2010, vgl. Abschnitt 3.2.2) zum Gebrauch generischer Nomina zur Einführung des Definiendums.

[104] sic!

31) Der Begriff der Priorität (lat . prior = der Vordere) bezeichnet im Allgemeinen den Vorrang einer Sache , eines Bewertungsmaßstabes oder einer Person gegenüber einer anderen .[105]

32) Der Begriff „ Nettostromerzeugung " bezeichnet dabei genauer die Differenz zwischen der insgesamt erzeugten Strommenge (Bruttostromerzeugung) und dem Eigenbedarf der verschiedenen Kraftwerke .[106]

Das Nomen „Form" erscheint dagegen in einer Vielzahl unterschiedlicher Verwendungskontexte und bezieht sich dabei sowohl auf äußerliche Aspekte von Gegenständen als auch auf Hyperonymie-/Hyponymie-Beziehungen, vgl. Beispiele 33) und 34):

33) Auch der sogenannte Netzparallelbetrieb ist eine Form des Verbundbetriebs.

34) Schwerkraftfundamente (Offshore) werden beispielsweise in der Form von großen Betongewichten auf dem Seeboden abgelegt, die so schwer und stabil sind, dass sie die Kräfte der WEA ohne weitere Verankerungen am Seeboden aufnehmen.

Die Partikel „dar" ist höchstwahrscheinlich ein abgetrennter Bestandteil des Verbs „darstellen", das zu den von Walter (2010) genannten Extraktionsmustern gehört (vgl. Tabelle 4). Analog besteht Grund zu der Annahme, dass das häufige Auftreten der Präpositionen „aus" und „zwischen" zumindest teilweise durch hohe Frequenzen entsprechender Verben (z. B. „bestehen" bzw. „unterscheiden") determiniert wird. Zudem zeigt der deutsche Goldstandard einen hohen Anteil von Relativsätzen, die durch ein Relativpronomen eingeleitet werden und in der Regel relevante Attribute oder Relationen des durch den Kontext erläuterten Begriffs zu einem anderen Begriff darstellen.

35) Das sogenannte Abreißflattern stellt ein weiteres Problem dar, bei welchem dynamischer Stall kurzzeitig über einem bestimmten Winkelbereich der Rotordrehung mit einer Struktureigenfrequenz auftritt.

36) Eine Verstärkung des Auftriebes wird durch die Erhöhung des so genannten „Anstellwinkels" bewirkt, welcher den Winkel der Querschnittlinie (Profilsehne) des Rotorblattes/ Tragfläche zu der anströmenden Luft beschreibt.

Die deutlich höhere Frequenz des Hilfsverbs „werden" in Verbindung mit dem bereits diskutierten häufigen Auftreten des Partizips II (*tag* VVPP) im Goldstandard ist ein Indiz für häufigen Passivgebrauch und damit ein typisches Merkmal allgemeiner Aussagen

37) Durch die sogenannte Leistungskennlinie kann die tatsächliche Leistungsabgabe einer Windkraftanlage angenähert werden.

38) Ein-, Zwei- oder Dreiblattrotoren werden als Schnelläufer bezeichnet.

[105] sic!
[106] sic!

Wie bereits im Falle des Russischen handelt es sich bei den meisten der darüber hinaus in der Tabelle zusammengetragenen Verben um Kandidaten für *knowledge patterns*. Sie sollen an anderer Stelle diskutiert werden. Analog zur bisherigen Darstellung fasst schließlich Tabelle 45 die deutschen Stopwörter zusammen.

Wort	**Absolute Frequenz Gold-standard (n=12073)**	**Absolute Frequenz NCL (n=9265)**	**Prüf-größe**	**p**	**KI (95%)**
Adverbien					
auch	59	68	4,9229	0,0265	[-0,0047, -0,0002]
dann	6	19	9,5272	0,0200	[-0,0027, -0,0005]
doch	0	12	13,4262	0,0002	[-0,0021, -0,0005]
einmal$_F$	0	5	-	0,0154	[0, 0,8372]
erst*	7	19	8,1501	0,0043	[-0,0026, -0,0004]
etwas	0	12	13,4262	0,0002	[-0,0021, -0,0005]
ganz	1	15	14,5233	0,0001	[-0,0025, -0,0006]
immer	3	17	12,4448	0,0004	[-0,0026, -0,0006]
nur*	13	25	6,8684	0,0088	[-0,0029, -0,0003]
jetzt*	0	12	13,4262	0,0002	[-0,0021, -0,0005]
mal$_F$	0	9	-	0,0005	[0, 0,3886]
recht$_F$	0	5	-	0,0154	[0, 0,8372]
schon	3	16	11,2713	0,0008	[-0,0025, -0,0005]
spät$_F$	0	5	-	0,0154	[0, 0,8372]
vielleicht$_F$	0	6	-	0,0067	[0, 0,6516]
wieder	2	10	6,2451	0,0125	[-0,0017, -0,0001]
Adjektive					
alle	8	20	7,8475	0,0051	[-0,0026, -0,0004]
bisherig$_F$	0	5	-	0,0154	[0, 0,8372]
eigen$_F$	0	5	-	0,0154	[0, 0,8372]
gut*	2	20	18,3276	< 0,0001	[-0,0031, -0,0009]
neu	3	15	10,1126	0,0015	[-0,0023, -0,0004]
Konjunktionen					
aber*	6	41	35,0409	< 0,0001	[-0,0054, -0,0024]
da	12	22	5,4427	0,0197	[-0,0026, -0,0001]
dass*	16	34	11,3427 (14,4627)	0,0008 (0,0001)	[-0,0038, -0,0009]
wenn	10	26	11,0306	0,0009	[-0,0033, -0,0007]
Nomen					
Frage$_F$	0	6	-	0,0067	[0, 0,6516]
Partikel					
ja	0	16	18,6237	< 0,0001	[-0,0027, -0,0008]
kein	14	24	5,2586	0,0218	[-0,0027, -0,0001]
nicht*	31	87	43,1361 (8,0207)	< 0,0001 (0,0046)	[-0,0091, -0,0046]
nichts$_F$	0	11	-	0,0001	[0, 0,3055]
selbst	2	10	6,2451	0,0125	[-0,0017, -0,0001]
Präpositionen					
an	55	79	12,6181	0,0004	[-0,0063, -0,0017]

für*	60	79	9,7058 (4,3776)	0,0018 (0,0364)	[-0,0059, -0,0012]
gegen	2	20	18,3276	< 0,0001	[-0,0031, -0,0009]
laut$_F$	0	5	-	0,0154	[0, 0,8372]
nach*	22	39	9,6586 (4,6221)	0,0019 (0,0316)	[-0,0040, -0,0008]
noch	14	29	9,1643	0,0025	[-0,0034, -0,0006]
Pronomen					
dies*	19	38	11,6409 (50,0263)	0,0006 (< 0,0001)	[-0,0041, -0,0010]
du$_F$	0	9	-	0,0005	[0, 0,3886]
er*	8	56	48,9867 (8,6647)	< 0,0001 (0,0032)	[-0,0071, -0,0036]
es*	25	80	44,7945 (6,2614)	< 0,0001 (0,0123)	[-0,0087, -0,0044]
ich*	0	68	86,5970 (14,9082)	< 0,0001 (0,0001)	[-0,0092, -0,0055]
ihr*	15	39	17,1240 (4,5897)	< 0,0001 (0,0322)	[-0,0045, -0,0014]
mehr	7	17	6,2750	0,0122	[-0,0023, -0,0002]
paar$_F$	0	8	-	0,0013	[0, 0,4493]
Sie|sie*	2	26	25,9134 (16,1172)	< 0,0001 (0,0001)	[-0,0038, -0,0014]
sie	24	42	10,2037	0,0014	[-0,0042, -0,0009]
unser$_F$	0	11	-	0,0001	[0, 0,3055]
wer$_F$	0	6	-	0,0067	[0, 0,6516]
wir	1	31	35,1281	< 0,0001	[-0,0045, -0,0020]
Verben					
brauchen$_F$	0	5	-	0,0154	[0, 0,8372]
bringen$_F$	2	9	-	0,0131	[0,0179, 0,8236]
erklären$_F$	0	6	-	0,0067	[0, 0,6516]
fahren$_F$	0	9	-	0,0005	[0, 0,3886]
geben	13	27	8,5021	0,0035	[-0,0032, -0,0005]
gehen	1	18	18,3476	< 0,0001	[-0,0029, -0,0009]
glauben$_F$	0	7	-	0,0029	[0, 0,5233]
haben	28	85	45,4720	< 0,0001	[-0,0091, -0,0046]
helfen$_F$	0	6	-	0,0067	[0, 0,6516]
kommen	7	21	10,1306	0,0015	[-0,0028, -0,0005]
machen	3	19	14,8280	0,0001	[-0,0029, -0,0007]
reden$_F$	0	5	-	0,0154	[0, 0,8372]
sagen	2	16	13,3646	0,0003	[-0,0025, -0,0006]
sehen	1	19	19,6286	< 0,0001	[-0,0030, -0,0009]
sollen	5	28	21,4321	< 0,0001	[-0,0039, -0,0013]
tun$_F$	0	9	-	0,0005	[0, 0,3886]
wollen	0	20	23,8317	< 0,0001	[-0,0032, -0,0011]

Tabelle 45: Kandidaten für deutsche Stopwörter

Neben den in der Tabelle dargestellten Stopwörtern erbringt der Vergleich des Goldstandards mit der Stichprobe aus den Extraktionskorpora Verteilungsunterschiede ebenfalls für die Lemmata „müssen“ (Prüfgröße 18,8196, $p < 0,0001$), „keine“ (Prüfgröße 20,1797, $p < 0,0001$), „jede“

(p = 0,0007), „einige“ (p = 0,0015), „wenige“ (p = 0,0015), „je“ (Prüfgröße 8,6326, p = 0,0033), „stark“ (Prüfgröße 8,0529, p = 0,0045), „viele“ (p = 0,0064), „folgend“ (p = 0,0064), „zeigen“ (Prüfgröße 7,2556, p = 0,0071), „allerdings“ (p = 0,0098), „gleich“ (p = 0,0098), „weiter“ (p = 0,0132), „mehrere“ (p = 0,0132), „beide“ (Prüfgröße 5,3731, p = 0,0204), „erreichen“ (Prüfgröße 5,0085, p = 0,0252), „bis“ (Prüfgröße 4,977, p = 0,0257), „arbeiten“ (Prüfgröße 3,8948, p = 0,0484), „pro“ (Prüfgröße 3,8948, p = 0,0484) und „liegen“ (Prüfgröße 3,8872, p = 0,0487). Auch hier ist auf die niedrigen Frequenzen und Effektstärken sowie die Notwendigkeit umfangreicherer Untersuchungen hinzuweisen. In Bezug auf das „Boostwort“ „jed“ und das „Stopwort“ „jede“ sowie eine Reihe weiterer Adjektive mit quantifizierender Bedeutung ist wiederum auf die Lemmatisierungsproblematik zu verweisen.

Im Ergebnis zeigt sich, dass der Goldstandard im Vergleich zu den Referenzdaten weniger quantifizierende („nur“) und temporale Adverbien („jetzt“), bewertende Adjektive („gut“), unterordnende und adversative Konjunktionen („dass“, „aber“), Negationen („nicht“) sowie unterschiedliche Arten von Pronomen einschließlich Personal-, Demonstrativ- und Possessivpronomen aufweist. Des Weiteren zeigen sich zumindest Tendenzen zur selteneren Verwendung von Modalverben („müssen“) sowie quantifizierender Adjektive („einige“, „wenige“, „viele“). Tendenziell können damit die Aussagen Walters (2010) sowie in Teilen auch Cramers (2011) bestätigt werden. Nicht zuletzt stützen auch diese Ergebnisse die bisherigen Ausführungen zu einem reduzierten Funktionalstil im Sinne positiver Sachaussagen und dem Überwiegen nebenordnender Strukturen.

5.2.2.4 Zwischenfazit

In Tabelle 39 war nach Wortartenverteilungen sowie lexikalischen Generalisierungssignalen (z. B. Frequenzadverbien), unpersönlichen Pronomen, generischen Nomina, definitionstypischen Adverbien sowie Stopwörtern wie den von Walter (2010) genannten Deixis, Definita, Demonstrativpronomen bzw. –adjektiven gefragt worden. Im Hinblick auf Wortartenverteilungen zeigt sich für das Russische, dass KRCs sich durch höhere Anteile von Nominalphrasen sowie geringere Anteile von Adverbien, pronominalen Adverbien (*tag* ADV-PRO, wobei mit diesem Tag Lemmata wie *как,* „wie“, *так,* „so“, *где,* „wo“, *иначе,* „anders“, und *здесь,* „hier“, markiert werden, so dass für russische KRCs von einer selteneren Verwendung von Deixis und Fragewörtern ausgegangen werden kann), Verben, Partikeln und Prädikativen auszeichnen. Diese Merkmale waren als Indikatoren für einen höheren Fachlichkeitsgrad wissenshaltiger Kontexte sowie einen um Objektivität, Faktenorientierung und Explizität bemühten Stil interpretiert worden. Die deutschen Ergebnisse stützen und differenzieren diese Annahmen. So zeichnet sich der deutsche Goldstandard insbesondere durch einen höheren Anteil des Partizips II aus, was als Hinweis auf den häufigen Gebrauch des als Generalisierungsmerkmal betrachteten Passivs interpretiert werden kann, jedoch kein zuverlässiges Indiz darstellt (das Partizip kann z. B. auch auf die Verwendung des Perfekts hinweisen, jedoch wurde auch das Hilfsverb „werden“ als „Boostwort“ identifiziert, womit in Verbindung mit dem Partizip II starke Hinweise auf einen häufigen Gebrauch des Passivs vorliegen). Des Weiteren zeigt der Vergleich von Wortartenfrequenzen Tendenzen zur häufigeren Verwendung nebenordnender Strukturen sowie zur selteneren Verwendung von Eigennamen, d. h. auf einzelne Individuen bezogenen Benennungen, was auf die Formulierung allgemeiner Aussagen hindeutet. Des Weiteren wurden niedrigere Anteile von Modalverben sowie attribuierender Indefinitpronomen (*tag* PIAT) und Possessivpronomen (*tag* PPOSAT) gemessen. Tendenziell bestätigen die Ergebnisse damit

auch die Aussagen Walters (2010) über Stopwörter in deutschen definitorischen Aussagen. Zwar können *tagging*-Fehler aufgrund der automatischen Wortarten-Annotation mit TreeTagger (Schmid 1994) nicht ausgeschlossen werden, jedoch wurden Vergleiche mit zwei unterschiedlichen Referenzdatensätzen – jeweils einem allgemeinsprachlichen (Deutsch) bzw. nur bedingt fachlichen (Russisch) Korpus sowie einer Zufallsstichprobe von Nicht-KRCs aus den jeweiligen Extraktionskorpora – durchgeführt, so dass für Aussagen über typische Merkmale von KRCs eine größere Datengrundlage herangezogen und allgemein „fachsprachliche“ Merkmale von solchen, die KRCs von generischen Aussagen unterschiedlichen Fachlichkeitsgrads unterscheiden, zumindest ansatzweise getrennt werden konnten. Die Angabe von Konfidenzintervallen für den Unterschied zwischen den Anteilswerten bzw. – für den Fisher-Test – für das Chancen-Verhältns (vgl. R Development Core Team 2012: 1138 f.) hilft zudem dabei einzuschätzen, welche der beobachteten Unterschiede z. B. für Extraktionsexperimente auch von praktischer Relevanz sein können.

In Bezug auf die Frequenzen von Lemmata (Boost- und Stopwörter) sind Verallgemeinerungen aufgrund der geringen Größe der untersuchten Datensätze mit entsprechend hohen Anteilen niedrigfrequenter Lemmata sowie geringen Verteilungsunterschieden kaum möglich. Darüber hinaus wirken sich Lemmatisierungsfehler bei geringen empirischen Frequenzen besonders negativ auf die statistischen Signifikanztests aus und können u. U. zu unterschiedlichen Lemmatisierungen und somit verzerrten Testergebnissen führen. Zwar konnten mit Hilfe statistischer Tests zahlreiche Kandidaten für russische Boostwörter gefunden werden, der Vergleich mit den Extraktionskorpora bestätigte jedoch nur einige wenige generische Nomina sowie eine Reihe von Kandidaten für *knowledge patterns*, überwiegend imperfektive Verben; andere in der Literatur beschriebene Typen von Boostwörtern konnten jedoch nicht gefunden werden. Hinsichtlich der identifizierten russischen Stopwörter kann konstatiert werden, dass diese die bereits bei der Analyse von Wortartenverteilungen beschriebenen Tendenzen bestätigen. So deuten Verteilungsunterschiede für bestimmte qualifizierend-bewertende Adjektive sowie quantifizierende Adverbien, Konjunktionen, Partikel und Pronomen sowie Verben an, dass die bisher beschriebenen Merkmale, z. B. der Verzicht auf eine persönliche Darstellung, auch auf lexikalischer Ebene Ausdruck finden, wobei jedoch angesichts der niedrigen, anhand der Konfidenzintervalle gemessenen Effektstärken fraglich ist, inwieweit Frequenzen lexikalischer Einheiten für praktische Extraktionsexperimente relevant sein können. Nicht zuletzt deutet das Überwiegen imperfektiver Verben bei gleichzeitigem Verzicht auf das Verb *быть („sein“)* auf die überwiegende Verwendung des Tempus‘ Präsens hin, womit ein wesentliches linguistisches Signal für die Allgemeinheit einer Aussage (vgl. Drescher 1992: 66) erfüllt ist.

Für das Deutsche ergab die Suche nach Boostwörtern neben Kandidaten für *knowledge patterns* und wenigen generischen Nomina auch Adjektive, mit deren Hilfe die zeitliche Gültigkeit von Informationen markiert wird, sowie Elemente, die Hinweise auf die Struktur wissenshaltiger Kontexte geben (Konjunktionen „entweder“, „oder“, Relativpronomen „welcher“). Damit konnten unterschiedliche Aussagen aus der analysierten Literatur auch für deutsche wissenshaltige Kontexte zumindest in der Tendenz bestätigt werden, wobei auch hier die oben beschriebenen Einschränkungen gelten. Allerdings finden sich auch unter den deutschen Boostwörtern keine Frequenzadverbien – einige allgemeinsprachliche Adverbien dieser Art, z. B. „immer“, finden sich vielmehr unter den Stopwörtern –, das unpersönliche Pronomen „man“ konnte ebenfalls nicht als Boostwort identifiziert werden, dasselbe gilt für definitionstypische bzw. generalisierende Adverbien. Tatsächlich besteht Grund zu der Annahme, dass zumindest die von Cramer (2011) angeführten

„definitionstypischen“ Adverbien sehr selten auftreten[107] und daher für eine Untersuchung dieser Lemmata deutlich größere Datensätze erforderlich wären. Generell scheint die hier beschriebene Untersuchung die Aussage Cramers (ebd.: 189), derzufolge „[…] viele lexikalische Merkmale bzgl. des Merkmalstyps recht häufig, aber in ihrer konkreten Merkmalsausprägung selten“ sind, zu stützen. Die Analyse von Stopwörtern stützt dagegen, wie bereits festgestellt, tendenziell die Aussagen Walters (2010) und auch Cramers (2011).

Im Ergebnis kann damit festgehalten werden, dass es sich zumindest bei einem Teil der in der gesichteten Literatur beschriebenen lexikalischen Signale für das Vorliegen eines wissenshaltigen Kontexts bzw. einer Definition um eher seltene Phänomene handelt, die durch quantitative Analysen nicht ohne Weiteres bzw. nur auf Grundlage deutlich größerer Datensätze bestätigt werden können, für praktische Extraktionsaufgaben aber von untergeordneter Bedeutung scheinen. In Bezug auf empirisch beobachtbare wissenshaltige Kontexte kann im Gegensatz zu notwendigerweise unvollständigen Beschreibungen in der Literatur von einer weitaus größeren Variationsbandbreite von Merkmalsausprägungen ausgegangen werden, wobei hinsichtlich der Relevanz einzelner Merkmale, d. h. dafür, wie zuverlässig bestimmte Merkmale tatsächlich das Vorliegen eines wissenshaltigen Kontexts signalisieren, bislang kaum Informationen vorliegen. Tendenziell bestätigen die bisherigen Analysen jedoch in Bezug auf einzelne lexikalische Merkmale wissenshaltiger Kontexte die aus der Literatur entnommenen Hypothesen für das Deutsche und deuten auch auf deren Übertragbarkeit auf das Russische hin. Für das Vorliegen eines deutlich abgrenzbaren Funktionalstils des Erklärens und Definierens wurden zwar Indizien gefunden, insgesamt scheinen die bisherigen Ergebnisse für das Postulat eines solchen Stils jedoch als zu schwach.

5.2.2.5 Bausteine für „knowledge patterns“

5.2.2.5.1 Russisch

In den 292 Sätzen des Goldstandards für das Sub-EKR wurden 375 Vorkommen potentieller Bausteine für wissenshaltige Kontexte identifiziert. Abbildung 16 stellt die am häufigsten vorkommenden (mindestens 5 Fundstellen) Elemente zusammen mit ihren absoluten Frequenzen überblicksartig dar.

Ein Vergleich mit Tabelle 42 zeigt zunächst, dass die Verben *состоять* (bestehen (aus)), *служить* (dienen (zu)), *предназначить* (vorsehen, bestimmen, vgl. Beispiele 24) und 25)), *представлять (собой)* (darstellen, vgl. Beispiel 20)), *обеспечивать* (gewährleisten), *устанавливать* (aufstellen, installieren), *использовать* (benutzen), *устанавливаться* (aufgestellt werden) und *применять* (verwenden) bereits als Boostwörter für den russischen Goldstandard identifiziert worden waren. Darüber hinaus handelt es sich bei den Verben *установить* (aufstellen, installieren)[108], *применяться* (verwendet werden) und *использоваться* (benutzt werden) um Aspektpartner bzw. lexikalisierte Passivformen bereits ermittelter Boostwörter. Die genannten Verben korrespondieren im Übrigen mit in der Literatur für andere Sprachen beschriebenen Extraktionsmustern (vgl. Tabelle 4) und sind mithin bereits bekannten semantischen Paradigmen

[107] Das Wortschatz-Portal der Universität Leipzig ordnet „normalerweise“ der Häufigkeitsklasse 11, „typischerweise“ der Häufigkeitsklasse 15, „gängig“ der Häufigkeitsklasse 16 und „üblicherweise“ der Häufigkeitsklasse 12 zu, d. h. das häufigste deutsche Wort „der“ ist 2^{16}-mal häufiger als „gängig“. http://wortschatz.uni-leipzig.de/abfrage/, http://wortschatz.uni-leipzig.de/html/faq/hkl.html, abgerufen am 17.1.2014.

[108] Allerdings war dieses Lemma als Stopwort für den Vergleich Gold-Extraktionskorpus identifiziert worden.

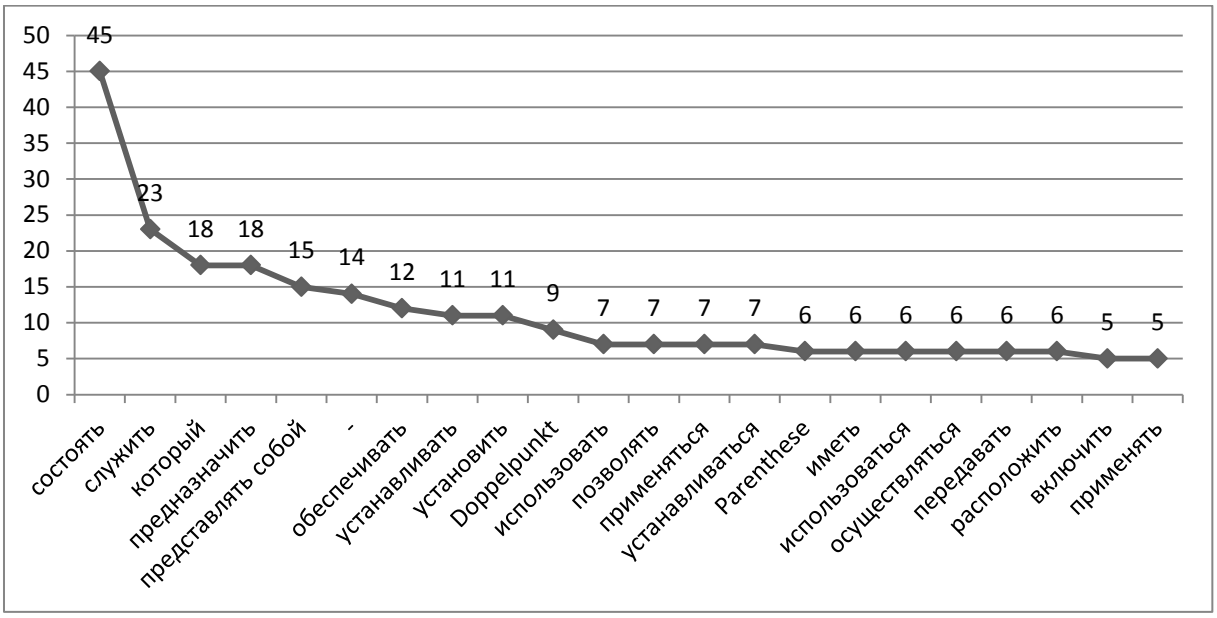


Abbildung 16: Bausteine für russische knowledge patterns in den KRCs des Sub-EKR

zuzuordnen. Neben diesen lexikalischen Bausteinen wurde auch das im Vergleich der Gold-Daten mit der Stichprobe aus dem Extraktionskorpus als Boostwort-Kandidat identifizierte Relativpronomen *который* (18 Fundstellen) bereits diskutiert (vgl. Beispiele 22) und 23)). Auch Relativsätze werden zudem in der Literatur als Merkmale wissenshaltiger Kontexte in anderen Sprachen genannt.

Der Gedankenstrich erscheint als Platzhalter für eine Kopula häufig (14 Fundstellen) in den wissenshaltigen Kontexten des Goldstandards für das Sub-EKR und verbindet in diesen in der Regel Definiendum und *genus proximum* (vgl. Beispiele 21) und 22)). Da Verwendungen des Gedankenstrichs als Kopula-Ersatz von konkurrierenden Verwendungsweisen nur schwer abgrenzbar sind und für den gesamten russischen Goldstandard lediglich Wortfrequenzen ermittelt wurden, liegen keine Informationen zur Verteilung dieses Elements im gesamten russischen Goldstandard im Gegensatz zu Vergleichsdaten vor.

Ein eher sekundärer Musterkandidat ist des Weiteren der Doppelpunkt (9 Fundstellen), der überwiegend in Verbindung mit anderen Musterkandidaten und in spezifischen Verwendungskontexten (z. B. im Zusammenhang mit Listen) auftritt, vgl. Beispiel 39).

39) Тепло в двигателях отводится двумя способами: жидкостью (жидкостная система охлаждения) или воздухом (воздушная система охлаждения).
[*Übersetzung:* Die entstehende Wärme wird auf zweierlei Arten aus dem Motor abgeleitet: durch Flüssigkeit (Flüssigkeitskühlung) oder durch Luft (Luftkühlung).]

Das Verb *позволять* (erlauben, 7 Fundstellen) erscheint überwiegend mit einem zweiten, variablen Verb im Infinitiv. In der Regel leitet *позволять* die Beschreibung von Funktionen des Definiendums ein, vgl. Beispiele 40) und 41).

40) Манометр позволяет проверять давление воздуха, как в воздушных баллонах, так и в тормозных камерах системы пневматического привода.
[*Übersetzung:* Das Manometer erlaubt die Messung des Luftdrucks sowohl in den Luftspeichern als auch in den Bremskammern des pneumatischen Antriebssystems.]

41) Стояночная тормозная система позволяет удерживать автомобиль неподвижно относительно дороги.
[*Übersetzung:* Die Standbremse erlaubt das Feststellen des Fahrzeugs auf der Straße.]

Rein quantitativ kann *позволять* jedoch nicht als Boostwort angesehen werden, da es weder im Vergleich des gesamten Goldstandards mit dem RIC (Prüfgröße: 3,2462, p = 0,0716) noch mit der Stichprobe generischer Sätze aus den Extraktionskorpora (Prüfgröße: 2,0538, p = 0,1518) signifikante Verteilungsunterschiede aufweist.
Parenthesen (6 Fundstellen) sind geklammerte Einschübe, die entweder ein Synonym oder als Konkretisierung nach einer längeren Erklärung das Definiendum selbst, mitunter auch eine stark vereinfachte Begriffserklärung enthalten, vgl. Beispiel 42).

42) Перфорация (дырки в диске), по заверениям производителей, способствует вентиляции дисков при эксплуатации автомобиля в предельных условиях.
[*Übersetzung:* Eine Perforation (Löcher in der Felge) unterstützt nach Aussagen der Hersteller die Belüftung der Felge unter Extrembedingungen.]

Das Verb *иметь* (haben, 6 Fundstellen) setzt in der Regel das Definiendum zu einem seiner Bestandteile in Beziehung und drückt damit eine Meronymie-Beziehung aus, vgl. Beispiele 43) und 44).

43) Впускной канал каждого цилиндра имеет по одной форсунке.
[*Übersetzung:* Der Einlasskanal jedes Zylinders hat eine Düse.]

44) Механизм переключения передач имеет три вилки переключения, входящие в соединение с подвижными элементами коробки передач.
[*Übersetzung:* Der Schaltmechanismus hat drei Schaltgabeln, die in die beweglichen Teile des Getriebes greifen.]

Für dieses Lexem kann jedoch aufgrund von 19 Fundstellen im gesamten russischen Goldstandard und 12 Fundstellen in der RIC-Stichprobe nicht konstatiert werden, dass das Verb für KRCs spezifisch sei (Prüfgröße Chi-Quadrat-Test < 1, p = 0,4559). Auch im Vergleich zu der Stichprobe generischer Sätze aus den Extraktionskorpora zeigt sich, dass das Verb in Bezug auf wissenshaltige Kontexte über kein Diskriminationspotential verfügt (21 Fundstellen im Extraktionskorpus, Prüfgröße Chi-Quadrat-Test < 1, p = 0,4899).
Das Verb *осуществляться* (umgesetzt werden, realisiert werden, 6 Fundstellen) setzt in den untersuchten KRCs als Kopula einen Begriff zu seiner Funktion oder seinem Verwendungszweck bzw. in Verbindung mit einer Präposition auch zu seiner Ursache oder aber zu prozessualen Angaben in Beziehung, vgl. Beispiele 45) und 46).

45) Теплоотвод в барабанных тормозных механизмах осуществляется через термопроводные колодочные накладки (см. раздел "Тормозные колодки"), массивную металлическую основу колодки и ребра охлаждения тормозного барабана.
[*Übersetzung*: Die Wärmeabfuhr wird in Trommelbremsen durch wärmeleitfähige Beläge auf den Bremsbacken (s. Abschnitt „Bremsbacken”), eine massive Metallbasis der Bremsbacken sowie Kühlrippen an der Trommel gewährleistet.]

46) В термосифонной системе охлаждения циркуляция осуществляется за счет разницы в плотности холодной и горячей жидкости.
[*Übersetzung:* Bei einer Thermosiphonkühlung beruht die Zirkulation auf dem Dichteunterschied zwischen der kalten und der heißen Flüssigkeit.]

Im gesamten russischen Goldstandard taucht das Verb insgesamt zehnmal auf, womit im Vergleich zum RIC (2 Fundstellen) ein statistisch signifikanter Verteilungsunterschied nicht festgestellt werden kann (Prüfgröße Chi-Quadrat-Test: 3,3005, p = 0,0693). Dieser Befund scheint auch dadurch bestätigt zu werden, dass die aktive Grundform des Verbs, *осуществлять*, mit zwei Fundstellen im RIC und neun Fundstellen im gesamten Goldstandard nicht als KRC-spezifisch angesehen werden kann. Auch im Vergleich zu den generischen Sätzen aus den Extraktionskorpora ist *осуществляться* nicht KRC-spezifisch (4 Fundstellen, Prüfgröße: 1,0552, p = 0,3043).
Das Verb *передавать* (übergeben, 7 Fundstellen) ist als eher sekundäres und möglicherweise fachgebietsspezifisches Muster zu charakterisieren, vgl. Beispiele 47) und 48).

47) Трансмиссионная сборка расположена слева от двигателя и передает развиваемый последним крутящий момент на передние колеса автомобиля.
[*Übersetzung:* Der Getriebeblock liegt links des Motors und überträgt das von diesem entwickelte Drehmoment auf die Vorderräder des Fahrzeugs.]

48) Карданная передача позволяет передавать крутящий момент между валами, расположенными под изменяющимся при движении автобуса углом.
[*Übersetzung*: Das Kardangelenk ermöglicht die Übertragung des Drehmoments zwischen Wellen, die sich bei Bewegung des Busses unter einem variablen Winkel schneiden.]

Im Vergleich des gesamten Goldstandards mit den Extraktionskorpora liegt das Verb über der Signifikanzgrenze von $p = 0{,}05$ (3 Fundstellen im Extraktionskorpus, Prüfgröße: 3,0567, p = 0,0804), erreicht aber im Vergleich zur RIC-Stichprobe signifikante Werte (1 Fundstelle im RIC, Prüfgröße: 6,5672, p = 0,0104), was die Annahme einer fachlichen Gebundenheit des Lexems stützt.
Das Verb *расположить* (anordnen, 6 Fundstellen) tritt ähnlich wie *предназначить* überwiegend als prädikativ gebrauchtes Adjektiv in Kurzform auf und bezeichnet die Position eines Objekts innerhalb eines komplexeren Gefüges, vgl. Beispiel 47). Auch dieses Verb zeigt im Vergleich zu Referenzdatensätzen keine signifikanten Verteilungsunterschiede (2 Fundstellen im RIC und p = 0,1156 für den Vergleich Gold-RIC und 2 Fundstellen im Extraktionskorpus, p = 0,1216 für den Vergleich Gold-Extraktionskorpora).
Das Verb *включить* (einschließen, beinhalten, 5 Fundstellen) erscheint ebenfalls überwiegend als prädikativ gebrauchtes Adjektiv in Kurzform und drückt eine Meronymie-Beziehung aus, vgl. Beispiele 49) und 50).

49) В гидравлический привод включены вакуумный усилитель и двухконтурный регулятор давления задних тормозов.
[*Übersetzung:* Der Hydraulikantrieb enthält einen Vakuumverstärker und einen zweikreisigen Druckregler für die hinteren Bremsen.]

50) Регулятор давления включен в оба контура тормозной системы, и через него тормозная жидкость поступает к обоим задним тормозным механизмам.
[*Übersetzung:* Ein Druckregler ist in beiden Bremskreisen enthalten und über ihn gelangt die Bremsflüssigkeit zu den beiden hinteren Bremsvorrichtungen.]

In den 560 Sätzen des gesamten russischen Goldstandards erscheinen insgesamt zwölf Formen von *включить* sowie seines imperfektiven Aspektpartners *включать*, in den 560 Referenzsätzen des RIC dagegen sechs Formen und in den Referenzsätzen aus den Extraktionskorpora sieben Formen. Anzumerken ist, dass auch hier Lemmatisierungsfehler im Ergebnis der automatischen Bearbeitung der Daten mit TreeTagger nicht ausgeschlossen werden können, so dass angesichts der in beiden Datensätzen niedrigen Frequenz des Lexems eine aggregierte Betrachtung für beide Aspektpartner und der Ableitungen von diesen sinnvoller erscheint als eine Einzelanalyse. Allerdings ergibt ein Signifikanztest für beide Vergleiche ein nicht-signifikantes Ergebnis (Prüfgröße 0,8576, p = 0,3544 für den Vergleich Gold-RIC und Prüfgröße 0,3082 und p = 0,5788 für den Vergleich Gold-Extraktionskorpora).

5.2.2.5.2 Deutsch

In den 337 Sätzen des Goldstandards für das Sub-EKD wurden 444 Vorkommen potentieller Bausteine für wissenshaltige Kontexte identifiziert. Abbildung 17 stellt die am häufigsten vorkommenden (mindestens 5 Fundstellen) Elemente zusammen mit ihren absoluten Frequenzen überblicksartig dar.
Auch ein Teil der in Abbildung 17 angeführten möglichen Musterbestandteile wurde bereits in Tabelle 44 zur Gruppe der deutschen Boostwörter zugeordnet, und zwar konkret die Einheiten „sogenannt“ (vgl. Beispiele 28) und 29)), „bestehen“, „bezeichnen“ (Beispiele 27), 31), 32) und 38)), „einsetzen“, „nennen“ und „unterscheiden“ (vgl. Beispiel 30)). Für eine Reihe anderer Einheiten konnten Verteilungsunterschiede dagegen nur für eines der beiden Vergleichskorpora nachgewiesen werden. Auch auf das häufige Auftreten von Relativsätzen war bereits verwiesen worden (vgl. Beispiele 35) und 36)). Zudem wurde auch die abgetrennte Partikel „dar“ (von „darstellen“, vgl. Beispiel 35)) als Boostwort ermittelt. Abbildung 17 zeigt des Weiteren, dass als mit Abstand häufigstes Prädikat in den untersuchten Sätzen mit dem Verb „sein“ (54 Fundstellen) eine sehr einfache Kopula auftritt, vgl. Beispiele 51) und 52).

51) Die Hauptaufgabe des Turms eines Horizontalachsenkonverters ist es nun, die Windenergienutzung in einer ausreichenden Höhe über Grund zu ermöglichen, sowie die statischen und dynamischen Belastungen des Rotors, des Triebstrangs und des Maschinenhauses aufzunehmen und sicher in den Untergrund abzuleiten.

52) Biomasse ist zusammen mit der Wasserkraft weltweit die wichtigste erneuerbare Energiequelle.

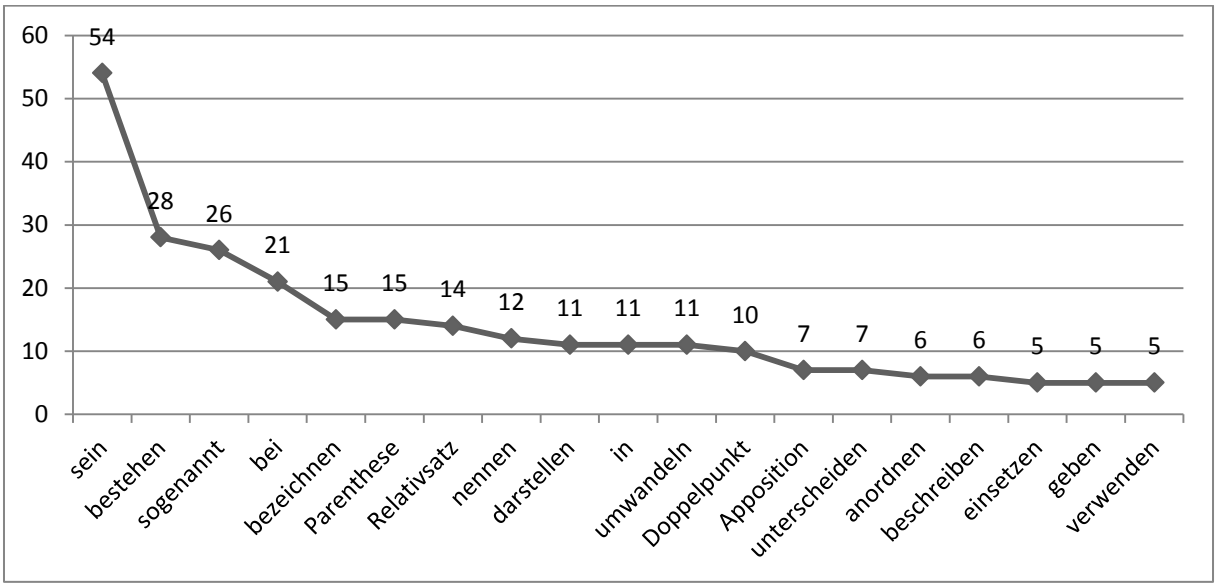


Abbildung 17: Bausteine für deutsche *knowledge patterns* in den KRCs des Sub-EKD

Für eine solch simple Kopula können kaum Unterschiede hinsichtlich der Verteilung im gesamten deutschen Goldstandard und den Vergleichsdaten angenommen werden (221 Fundstellen im NCL, Prüfgröße Chi-Quadrat-Test: 0,9212, p = 0,3372 für den Vergleich Gold-NCL sowie 242 Fundstellen, Prüfgröße: 0,0808 und p = 0,7762 für den Vergleich Gold-Extraktionskorpora). Das Verb „sein“ ist folglich nicht als für wissenshaltige Kontexte spezifisch anzusehen.
Das Verb „bestehen (aus)“ (28 Fundstellen) drückt häufig wie sein russisches Gegenstück *состоять (из)* eine zwischen zwei Begriffen bestehende Meronymie-Beziehung aus. Allerdings kann mit Hilfe von „bestehen“ auch das Material, aus dem ein Gegenstand besteht, benannt oder seine Bauform beschrieben werden. Diese Verwendungsweise des Verbs entspricht der in Tabelle 3 in Abschnitt 2.3.3.2 definierten semantischen Relation „Ursprung“.

53) Ein Darrieus-Rotor <u>besteht aus</u> mehreren (zwei bis vier) senkrecht gekrümmten Blättern, die sich gemeinsam mit dem Wind um eine vertikale Achse drehen.

54) Moderne Rotorblätter <u>bestehen aus</u> glasfaserverstärktem Kunststoff und werden in Halbschalen-Sandwichbauweise mit Versteifungsholmen oder -stegen im Inneren hergestellt.

Die Präposition „bei“ (21 Fundstellen) leitet üblicherweise Gültigkeitsbedingungen für die im KRC gegebene Aussage ein bzw. benennt das Fachgebiet oder den fachlichen Zusammenhang bzw. führt das Definiendum ein, vgl. Beispiele 55) und 56).

55) <u>Bei</u> der Energieumwandlung unter Verwendung einer Wärmekraftmaschine wird stets Wärme in (zunächst) mechanische Arbeit umgewandelt.

56) <u>Bei</u> der Anbindung einer Windkraftanlage an das Netz der öffentlichen Versorgung oder ein beliebiges Inselnetz wird unterschieden zwischen einer direkten und indirekten Netzkopplung.

Eine ähnliche Verwendungsweise der Präposition wird von Walter (2010: 49) beschrieben, Cramer (2011) bildet mit dieser Präposition eigene Extraktionsmuster. Auch Zifonun et al. (1997: 602) führen „bei“ als Beispiel für eine „Geltungsrestriktion“ an. Tatsächlich zeigt der Goldstandard im Vergleich zum NCL einen deutlich höheren Anteil dieser Präposition an der Gesamtmenge der Lemmata (37 Fundstellen, Prüfgröße: 21,9057, $p < 0{,}0001$), im Vergleich zum Extraktionskorpus kann dieser Unterschied jedoch nicht bestätigt werden, da der Anteil dieser Präposition im Goldstandard sogar leicht geringer ist als in der Stichprobe aus den Extraktionskorpora (114 Fundstellen, Prüfgröße: 0,0780, $p = 0{,}7800$). Wie bereits für das Russische können auch im deutschen Goldstandard Parenthesen (15 Fundstellen) als Signale für Begriffserklärungen interpretiert werden, vgl. Beispiele 57) und 58).

57) Die Spitzenverluste (oder Tip-Verluste) sind aerodynamische Verluste, die durch die Umströmung der Blattspitze von der Unterseite zur Oberseite des Blattes entstehen.

58) Man unterscheidet, ob sich der Rotor auf der dem Wind zugewandten Seite (Luvläufer) oder auf der dem Wind abgewandten Seite (Leeläufer) des Turmes befindet.

Das Verb „nennen“ (12 Fundstellen) korrespondiert mit den in der Literatur für eine Reihe unterschiedlicher Sprachen beschriebenen Mustern und auch mit den für das Russische ermittelten Boostwörtern und Bausteinen für *knowledge patterns*. Das Verb verbindet Definiendum und Definiens oder führt Synonyme oder Kurzerklärungen ein, vgl. Beispiele 59) und 60).

59) Der Wert, der die dem Wind entnommene Leistung ins Verhältnis mit der im Wind enthaltenen Leistung setzt, wird Betzscher Leistungsbeiwert () genannt und wurde von Albert Betz im Jahre 1926 ermittelt (siehe: Betzsches Gesetz).[110]

60) Die Folge waren weithin sichtbare periodische Reflexionen des Sonnenlichtes bei schönem Wetter und laufender Anlage, die Discoeffekt genannt wurden.

Mit der Präposition „in“ (11 Fundstellen, wobei auch Portmanteau-Formen, die aus Artikel und Präposition gebildet werden, Berücksichtigung finden) haben wir schließlich ein weiteres nicht-prädikatives Element vor uns, das anders als bereits genannte andere Präpositionen (vgl. die Boostwort-Tabelle in Abschnitt 5.2.2.3.2) kein direkter Bestandteil einer Begriffserklärung bzw. eines *knowledge patterns* ist, sondern vielmehr das Definiendum zu einer Begriffserklärung in Beziehung setzt, indem es – formal ähnlich wie eine Gültigkeitsangabe bzw. Geltungsrestriktion – das Definiendum als Geltungsbereich der Aussage einführt, vgl. Beispiele 61) und 62)[111].

61) In einer Brennstoffzelle verbindet sich der Wasserstoff mit dem Sauerstoff der Luft, um Strom zu erzeugen.

62) In einer Brennstoffzelle läuft eine sog. kalte Verbrennung ab, bei der aus einer exothermen Redox-Reaktion ohne den Umweg der Wärmeerzeugung elektrischer Strom produziert wird.

[110] sic!

[111] In Abschnitt 5.2.3.3.2 werden Präpositionalphrasen (z. B. auf „bei“ oder „in“), die das Definiendum enthalten, aber nicht Teil des Valenzrahmens eines Definitionsprädikats sind, als Situativkomplemente und damit als besondere Formen der Thematisierung des Definiendums klassifiziert.

In beiden Vergleichen des Goldstandards mit Referenzsätzen aus den Extraktionskorpora bzw. dem NCL wurden für „in“ jedoch keine signifikanten Verteilungsunterschiede ermittelt (244 Fundstellen im NCL, Prüfgröße: 0,1105, p = 0,7396 für den Vergleich Gold-NCL und 305 Fundstellen, Prüfgröße: 0,2665 und p = 0,6057 für den Vergleich Gold-Extraktionskorpus).
Bei dem Verb „umwandeln“ (11 Fundstellen) scheint es sich um ein fachgebietsspezifisches Lexem zu handeln, mit dessen Hilfe die wesentliche Funktion einer eng begrenzten Gruppe von Definienda beschrieben wird, vgl. Beispiele 63) und 64).

63) Eine Windenergieanlage (WEA) wandelt Windenergie in elektrische Energie um und speist diese zumeist in das öffentliche Stromnetz ein.

64) Über die Rotorblätter wird die Windenergie in Rotationsenergie umgewandelt und an den Generator weitergeleitet.

Das Verb war in Tabelle 44 aufgrund seiner fachlichen Spezifik nicht als Boostwort geführt worden, zeigt allerdings im Vergleich zu den Referenzdaten deutliche Verteilungsunterschiede und kann mithin als für den Goldstandard typisches Prädikat angesehen werden (0 Fundstellen im NCL, Prüfgröße: 10,5828, p = 0,0011 für den Vergleich Gold-NCL und 5 Fundstellen, Prüfgröße: 4,2423 und p = 0,0394 für den Vergleich Gold-Extraktionskorpus).
Der Doppelpunkt (10 Fundstellen) verbindet als nicht-sprachliches Element ein Definiendum mit nachfolgenden erklärenden Satzteilen und liegt formal bereits im Grenzbereich der durch die in Kapitel 2 entwickelte Definition und die in Kapitel 4 beschriebenen Validitätskriterien abgegrenzten Kategorie des wissenshaltigen Kontexts. So sind in Beispiel 65) Definiendum und Definiens sprachlich unabhängig voneinander – je nach kommunikativem Zusammenhang hätte eine äquivalente Aussage beispielsweise auch durch eine Überschrift mit nachfolgender Begriffserklärung in einem eigenen Satz realisiert werden können. Des Weiteren ähnelt ein solcher Kontext strukturell dem von Pearson (1998: 162 ff.) beschriebenen Typus der *dictionary type definition*, in der die Verbindung zwischen Lemma und Definition durch nicht-sprachliche Elemente hergestellt wird. Beispiel 66) ist indes anders als die während der Erstellung des Goldstandards als ungültige wörterbuchähnliche KRC-Kandidaten aus dem Goldstandard entfernten Sätze weder ungrammatisch noch stark vereinfachend und im Sub-EKD in einen normalen Textzusammenhang eingebunden. Aufgrund dieser Eigenschaften verblieb Beispiel 66) im Goldstandard, markiert dabei aber, wie festgestellt, einen Grenzfall.

65) Eine weitere Möglichkeit die Anlagen nach dem Wind auszurichten sind die so genannten passiven Systeme: Windnachführung durch selbstständigen Nachlauf von Leeläufern und Windfahnen für Luvläufer.[112]

66) Stall-Regelung (aerodynamischer Abriss): Steigt die Windgeschwindigkeit über ein bestimmtes Maß hinaus, reißt durch die spezielle Flügelform die Luftströmung AUF der Saugseite des Rotorblattes an der Hinterkante beginnend ab und begrenzt so die Drehzahl.

Appositionen als syntaktische Realisierungsformen von Begriffserklärungen treten insgesamt siebenmal in den untersuchten Goldstandard-Sätzen auf. Das Definiendum kann dabei ähnlich wie in

[112] sic!

Parenthesen entweder Element der Apposition (Beispiel 67)) oder aber dieser als Bezugswort vorangestellt sein (Beispiel 68)).

67) Bei einer Betrachtung der vertikalen Unterteilung der Atmosphäre ist alleine deren untere Schicht, die Troposphäre, für eine Nutzung der Windenergie von Interesse.

68) In der Gondel, dem Maschinenhaus eines riesigen Windkraftrads, wird die Drehbewegung der Rotorblätter aufgenommen und über Induktionsspulen in der Turbine in fließenden Strom verwandelt.

Der Vollständigkeit halber anzumerken ist, dass – obwohl Appositionen nicht auf lexikalischer, sondern vielmehr auf syntaktischer Ebene zu beschreiben sind – dieses Element bereits hier angeführt wird, da es im Zusammenhang mit der Untersuchung der Bausteine für *knowledge patterns* leicht zu erheben war.
Das Verb „anordnen“ (6 Fundstellen) tritt in den untersuchten Sätzen überwiegend als Zustandspassiv auf und beschreibt die Position des Definiendums relativ zu einem anderen Teil eines komplexen Ganzen oder die Art seiner Anordnung, vgl. Beispiele 69) und 70).

69) Die zylindrische Säule des Turmes ist auf dem Dach des Maschinenraumes angeordnet.

70) Die vertikalen Einleitflächenelemente sind radial angeordnet, wodurch die Windkraftanlage windrichtungsunabhängig ist.

Im gesamten Goldstandard treten Formen des Verbs insgesamt zehnmal auf (4 Fundstellen für „anordnen“ und 6 Fundstellen für „angeordnet“), in den zum Vergleich herangezogenen Sätzen des NCL jedoch kein einziges Mal. Insgesamt kann das Verb damit als für den Goldstandard eher typisch gelten (0 Fundstellen, $p = 0{,}0067$ für den Vergleich Gold-NCL und $p = 0{,}0020$ für den Vergleich Gold-Extraktionskorpus).
Das Verb „beschreiben“ (6 Fundstellen) bezieht sich – zumindest in den im Goldstandard identifizierten Beispielen – überwiegend auf numerische bzw. mathematisch-quantitative Angaben (vgl. Beispiele 71) und 72)) und scheint damit auf spezifische Verwendungskontexte festgelegt.

71) Die Energierücklaufzeit beschreibt die Zeit, die vergeht, bis ein Kraftwerk genauso viel Energie erzeugt hat, wie zu seiner Produktion, Transport, Errichtung, Betrieb usw. benötigt wurde.

72) Der sogenannte Abstandsfaktor kA beschreibt den zwischen zwei benachbarten Konvertern minimal einzuhaltenden Abstand.

Im Vergleich zum NCL zeigt der Goldstandard für dieses Verb signifikante Verteilungsunterschiede (0 Fundstellen, Prüfgröße: 7,5306, $p = 0{,}0061$), die allerdings im Vergleich mit den Extraktionskorpora nicht bestätigt werden können (4 Fundstellen, Prüfgröße: 2,6987, $p = 0{,}1004$). Das Verb „einsetzen“ (5 Fundstellen) dagegen wird für die Angabe von Verwendungskontexten und Funktionen genutzt, vgl. Beispiele 73) und 74).

73) Der Savonius-Rotor wird insbesondere für den mechanischen Antrieb von Wasserpumpen gelegentlich eingesetzt.

74) Insbesondere bei großen Windkraftanlagen wird aufgrund der diskutierten Nachteile der Stall-Regelung die Verstellung des Blattwinkels (Pitchregelung) zur Regelung der dem Wind entnommenen Leistung eingesetzt.

Das Verb „geben“ (5 Fundstellen) wird in unpersönlichen Sätzen zum Ausdruck von Hyponymie/Hyperonymie-Beziehungen in Form klassischer Kategorisierungen oder aber extensionaler Definitionen verwendet, vgl. Beispiele 75) und 76).

75) Es gibt im Wesentlichen zwei Bauformen von Windkraftanlagen: solche mit senkrechter Drehachse und solche mit waagrechter Achse.

76) Weltweit gibt es viele verschiedene Winde und Windsysteme, wie zum Beispiel den Passat, Monsun, Föhn, den Mistral, die Bora oder den Scirocco.

Als Boostwort kann das Verb aufgrund der statistischen Analyse allerdings nicht betrachtet werden, da es im Vergleich zum NCL im Goldstandard signifikant seltener erscheint (27 Fundstellen, Prüfgröße: 8,5021, p = 0,0035), im Vergleich zur Stichprobe aus den Extraktionskorpora mit dem gesamten Goldstandard jedoch annähernd gleiche Verteilungen aufweist (15 Fundstellen, Prüfgröße: 0,1107, p = 0,7394). Nicht zuletzt dient auch das Verb „verwenden“ (5 Fundstellen) zur Angabe von Funktionen bzw. Verwendungskontexten eines Begriffs, vgl. Beispiele 77) und 78).

77) Das Schalenkreuzanemometer wird für Windgeschwindigkeitsmeßgeräte verwendet (siehe Kapitel 4.3.1).

78) Es werden prinzipiell drei verschiedene Arten von Generatoren zur Stromerzeugung verwendet: der Asynchrongenerator, der Synchrongenerator und der Gleichstromgenerator.

Auch für dieses Verb gilt, dass es zwar im Vergleich zum NCL im Goldstandard deutlich häufiger erscheint (2 Fundstellen, Prüfgröße: 7,7906, p = 0,0053), dieser Effekt im Vergleich zu den Extraktionskorpora jedoch nicht bestätigt werden kann (9 Fundstellen, Prüfgröße: 1,9603, p = 0,1615).

5.2.2.5.3 Semantische Typen von „knowledge patterns“

a) Prozesstypen und Verbklassen

Im Merkmalskatalog war auch die Frage nach den Prozesstypen bzw. den semantischen Klassen derjenigen Verben, die als Bestandteile von *knowledge patterns* fungieren, aufgeworfen sowie danach gefragt worden, ob es sich bei diesen Verben um „allgemeine“ oder sogar „semantisch leere“ Verben handelt. Neben diesen beiden Fragen ist zudem von Interesse, inwieweit sich die Bausteine von *knowledge patterns* auch hinsichtlich anderer Merkmale (z. B. Variationsbreite, durchschnittliche Frequenzen im untersuchten Satzkorpus etc.) von den Verben in solchen Sätzen unterscheiden, denen Generizität unterstellt werden kann (vereinfacht ausgedrückt: Nicht-KRCs). Diese Fragen sollen nun näher untersucht werden, um eine genauere Vorstellung davon zu erlangen, welche Arten

von Verben zu Zwecken der Begriffsexplikation bzw. –exposition genutzt werden und wie dies geschieht.

Hinsichtlich der Frage nach semantischen Typen bzw. Verbklassen von Verben in *knowledge patterns* war auf die Prozesstypen der SFL bereits eingangs kurz verwiesen worden. Neben diesem insbesondere für die Beschreibung des Englischen angewandten Ansatz existieren sowohl für das Deutsche als auch das Russische eigene semantische Typologien für Verben bzw. entsprechende Klassifikationsschemata. Anzumerken ist in diesem Zusammenhang allerdings, dass der Begriff des „semantischen Typs“ bzw. der „Verbklasse“ nicht klar umrissen ist. Zwar werden lexikalische Einheiten etwa in Thesauri[113] traditionell nach semantischen Kriterien zusammengefasst, einheitliche Typologien semantischer Klassen etwa von Verben existieren nach gegenwärtigem Kenntnisstand selbst innerhalb von Sprachgemeinschaften jedoch nicht. Beispielsweise präsentieren Schulte im Walde und Brew (2002: 224 f.) 14 Verbklassen für eine Gruppe von 57 deutschen Verben, die aufgrund des Vorliegens gemeinsamer semantischer, aber auch syntaktischer Eigenschaften gebildet wurden. Die Autoren weisen darauf hin, dass diese Klassen mit den Klassen nach Schumacher (1986) kompatibel seien, der die in seinem Wörterbuch aufgelisteten Verben nach onomasiologischen Kriterien sieben Makrofeldern zuordnet (ebd.: 12 ff.), und zwar:

- Verben der allgemeinen Existenz,
- Verben der speziellen Existenz,
- Verben der Differenz,
- Verben der Relation und des geistigen Handelns
- Verben des Handlungsspielraums,
- Verben des sprachlichen Ausdrucks,
- Verben der vitalen Bedürfnisse.

Im Artikel von Schulte im Walde und Brew (2002: 225) lauten die Verbklassen dagegen „Aspect“, „Propositional Attitude“, „Transfer of Possession (Obtaining)“, „Manner of Motion“, „Emotion“, „Announcement“, „Description“, „Insistence“, „Position“, „Support“, „Opening“, „Consumption“, „Weather“, was deutlich macht, dass die von Autoren vorgeschlagene semantische Klassifizierung deutscher Verben auf einer wesentlich feinkörnigeren Granularitätsstufe ansetzt.

Eine wiederum anderslautende Klassifizierung semantischer Typen für deutsche Verben implementiert die elektronische Sprachressource GermaNet[114] (Hamp/Feldweg 1997). Hierbei handelt es sich um ein sogenanntes lexikalisch-semantisches Netz nach dem Vorbild des „Princeton WordNet“[115] (Fellbaum 1998), in dem lexikalische Einheiten ebenfalls nach onomasiologischen Kriterien zu sogenannten *synsets* zusammengefasst sind. Für deutsche Verben enthält GermaNet umfangreiche Informationen. So ordnet es jedes Verb mindestens einer der folgenden semantischen Klassen zu:

- Besitz,
- Lokation,
- Gefühl,
- Körperfunktion,
- Gesellschaft,
- Kognition,
- Kommunikation,

[113] Für das Deutsche dürfte Dornseiff (2004[8]) als bekanntester Thesaurus gelten.
[114] http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml, abgerufen am 07.04.2014.
[115] http://wordnet.princeton.edu/, abgerufen am 07.04.2014.

- Konkurrenz,
- Kontakt,
- Natürliches Phänomen,
- Schöpfung,
- Veränderung,
- Verbrauch,
- Perzeption,
- Allgemein.

Auch diese Klassifikation unterscheidet sich von derjenigen Schumachers (1986) dadurch, dass sie feinkörnigere Klassen bildet, jedoch nimmt sie – obwohl sie von der durch Schulte im Walde und Brew (2002) vorgeschlagenen Klassifikation ebenfalls abweicht – ebenso wie diese Bezug auf die Arbeit Levins (1993) zur Klassifikation englischer Verben.

Für das Russische existieren mit dem Thesaurus von Baranov (1995)[116], dem semantischen Wörterbuch von Švedova (2007) sowie der Datenbank Lexicograf (Kustova/Padučeva 1994) Ressourcen, die zwar keine Beschreibung russischer Verben liefern, die mit GermaNet oder anderen deutschen Ressourcen vergleichbar wäre, jedoch für eine semantische Klassifizierung russischer Verben relevante Hinweise liefern können. Während der Baranow-Thesaurus in seiner online verfügbaren Fassung das russische Lexikon in sechs Gruppen[117] (группы), 20 Untergruppen (подгруппы) und 70 Abschnitte (отделы) gliedert, legen Kustova und Padučeva mit Lexicograf eine linguistisch detailliert ausgearbeitete Klassifikation allein für russische Verben vor. Die in der Lexicograf-Datenbank enthaltenen Verben werden den taxonomischen Kategorien Zustand, Prozess, Handlung und Ereignis zugeordnet. Des Weiteren werden aspektuelle Eigenschaften und solche der semantisch-syntaktischen Oberflächenstruktur (semantische Rollen, syntaktischer Frame) beschrieben. Švedova (2007) wiederum gruppiert die russischen Verben nach lexikalischen und syntaktischen Gesichtspunkten in hierarchisch organisierte Gruppen. Konkret differenziert die Autorin (ebd.) auf einer Makroebene zwischen semantisch „leeren“ Verben (глаголы с ослабленной знаменательностью, неполнознаменательные, Verben mit reduziertem Bedeutungsgehalt), zu denen Kopula-, Phasen-, Modal- sowie Relations- und Benennungsverben zählen, Verben der Existenz (быйтийные глаголы, Verben des Seins) und Handlungsverben (глаголы со значением активного действия, Verben des aktiven Handelns), welche die weitaus größte und am stärksten differenzierte Gruppe des Wörterbuchs ausmachen.

Während für die semantische Analyse von Prädikaten Verben in semantischen Ressourcen nachzuschlagen und die Ergebnisse dieses Prozesses auszuwerten sind, können numerische Merkmale der Verben dagegen mit einfachen Mitteln der deskriptiven Statistik untersucht und den Merkmalen der Verben in den zufällig ausgewählten Sätzen aus den Extraktionskorpora gegenübergestellt werden. Von Interesse ist dabei beispielsweise, in welchem Maße es Überschneidungen zwischen den beiden Datensätzen gibt und welcher der beiden Datensätze eine höhere Variation aufweist.

[116] Eine aktuellere Fassung des Thesaurus‘ ist online unter http://www.speakrus.ru/dict/index.htm zugänglich, Zugriff am 08.04.2014.

[117] Die Gruppen lauten: Welt der grundlegenden Begriffe (мир сущностей), Attribute der körperlichen Welt (атрибуты материального мира), Natur (природа), Mensch (человек), Gesellschaft (общество), Kultur (культура).

b) Methode

Bei der Auswahl einer Ressource für die Zuordnung von Verben zu semantischen Klassen ist neben der generellen Verfügbarkeit von Ressourcen sowie deren theoretischer Grundlage auch deren Abdeckung (Anzahl erfasster Verben) sowie die Frage, ob der Datenzugriff manuell oder automatisch vonstattengeht, zu beachten, da beispielsweise bei automatischem Zugriff und automatisierter Annotation mit einer höheren Konsistenz der Zuordnung zu rechnen ist. Angesichts der Vielzahl von Typologien für das Deutsche, die sich trotz Überschneidungen hinsichtlich ihrer theoretischen Grundlegung dennoch voneinander unterscheiden, scheint die Auswahl eines Ansatzes auf Grundlage theoretischer Überlegungen mehr oder weniger arbiträr. Hinsichtlich der Abdeckung ist GermaNet (Hamp/Feldweg 1997) der Arbeit von Schumacher (1986) jedoch deutlich überlegen: Enthält Schumachers Wörterbuch „über 1000 Verben“ (ebd.: 1), umfasste GermaNet bereits bei seiner Erstveröffentlichung „some 16.000 words“ (Hamp/Feldweg 1997) und deckt gegenwärtig (Version 8.0) 111.361 lexikalische Einheiten ab[118], Kunze und Lemnitzer (2007: 143) nennen die Zahl von 9000 Verben, die in GermaNet enthalten sind. Darüber hinaus stehen für GermaNet mit eigenen Perl- und Java-APIs Werkzeuge für den automatischen Datenzugriff, mit denen beispielsweise für jedes Verb Verklassen automatisch ausgelesen werden können, zur Verfügung[119].

Um Vergleichsdaten zu generieren, werden zunächst manuell in 337 zufällig ausgewählten generischen Vergleichssätzen (Nicht-KRCs) aus dem Sub-EKD alle einfachen Prädikate (d. h. Prädikate, die nicht Teil von Funktionsverbgefügen sind) extrahiert, wobei wie im Falle der Annotation von *knowledge patterns* Hilfs- und Modalverben nicht berücksichtigt werden. Auf diese Weise wird ein Satz von 362 Vergleichsverben generiert. Alsdann wird im Goldstandard für das Sub-EKD für diejenigen Elemente von *knowledge patterns*, die einfache Verben darstellen (d. h. es handelt sich um prädikativ verwendete Verben im Gegensatz zu anderen Wortarten, nicht-sprachlichen Elementen oder aber Funktionsverbgefügen), mit Hilfe der erwähnten Java-API eine Liste semantischer Klassen extrahiert. Insgesamt werden so für den Goldstandard semantische Klassen für 296 Verb-Okkurrenzen, für die Referenzdaten dagegen Klassen für 340 Verb-Okkurrenzen[120] aus GermaNet (Version 7.0) ausgelesen. Da allerdings die meisten Verben bzw. Wortformen in GermaNet mehr als einem Synset zugeordnet werden können – beispielsweise zeigt Abbildung 18 anhand einer graphischen Benutzeroberfläche für GermaNet, GermaNet Explorer, dass das Verb „einsetzen“ in acht unterschiedlichen Synsets erscheint und in Abhängigkeit von der in diesem Synset modellierten Bedeutung unterschiedliche Hyperonyme aufweist (z. B. „beginnen“ und „gebrauchen“) –, ist eine automatische Auswahl der korrekten semantischen Klasse kaum möglich. Diese wurde daher manuell unter Berücksichtigung der korrekten Bedeutung des betreffenden Verbs vorgenommen.

[118] http://www.sfs.uni-tuebingen.de/GermaNet/, abgerufen am 20.1.2014.

[119] http://www.sfs.uni-tuebingen.de/GermaNet/tools.shtml, abgerufen am 07.04.2014.

[120] Nicht für alle Verben können in GermaNet passende semantische Klassen gefunden werden, so dass sich die Zahl der im Korpus annotierten Verb-Okkurrenzen von der Zahl der ausgelesenen semantischen Klassen unterscheidet.

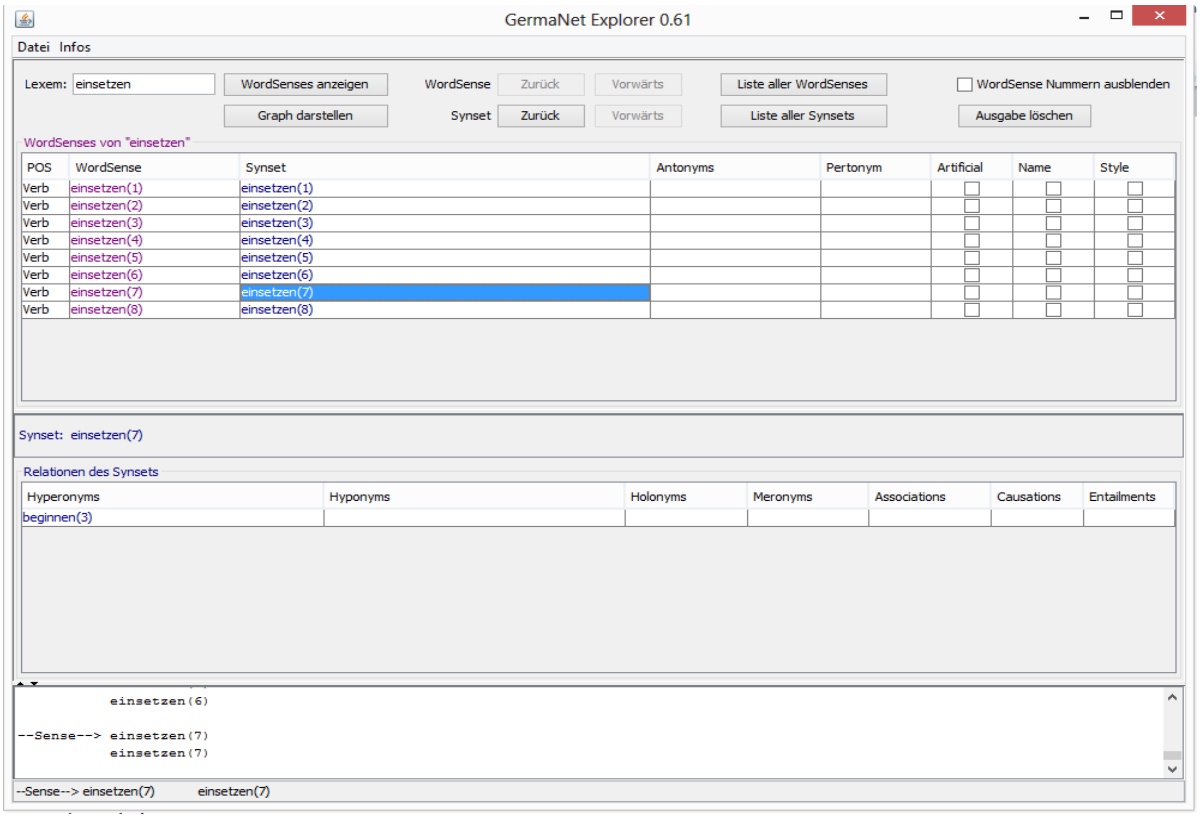


Abbildung 18: Screenshot aus dem GermaNet Explorer für das Verb „einsetzen“

Für das Russische werden manuell aus dem Wörterbuch von Švedova (2007) semantische Klassen für 230 Verb-Okkurrenzen aus der Gruppe der *knowledge patterns* des Sub-EKR extrahiert, wobei die Zuordnung zu semantischen Klassen zur Vermeidung allzu feingliedriger Unterscheidungen auf einem relativ hohen hierarchischen, aber dennoch semantisch aussagekräftigen Niveau erfolgt. Aspektpartner bzw. Formen, die durch den Suffix *–ся* als intransitiv oder passiviert markiert sind, werden dabei ihren unmarkierten Komplementärformen zugeordnet, auch wenn sie selbst nicht im Wörterbuch enthalten sind, um so die Abdeckung zu verbessern. Auch hier ist in der Regel eine fallbezogene Auswahl aus mehreren Kategorisierungsmöglichkeiten zu treffen. Ein Vergleichsdatensatz wird ebenso wie für das Deutsche erstellt, indem in 292 zufällig ausgewählten Nicht-KRCs aus dem Sub-EKR wiederum 284 einfache Prädikate markiert und im Wörterbuch nachgeschlagen werden. Dieser Prozess ergibt semantische Klassen für 199 Verb-Okkurrenzen.
Die auf diese Weise gewonnenen Daten zur Häufigkeit semantischer Klassen in den untersuchten Sätzen können wiederum statistisch mittels der bereits beschriebenen Hypothesentests ausgewertet werden. Die numerischen Merkmale der Verben aus *knowledge patterns* im Vergleich zu den Referenzdaten werden zudem mittels einfacher Perl-Skripte sowie in R mit Hilfe des Statistikpakets zipfR[121] (Evert/Baroni 2007) erhoben.

c) Ergebnisse
Tabelle 46 stellt analog zur bisherigen Darstellung von Berechnungen und Tests die Ergebnisse für das Deutsche im Überblick dar. Konkret wird dargestellt, wieviele Verb-Okkurrenzen jede

[121] http://zipfr.r-forge.r-project.org/. Eine Einführung in dieses Paket findet sich hier: http://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf, Zugriff am 08.04.2014.

semantische Klasse in den untersuchten Datensätzen auf sich vereint. Relative Frequenzen ergeben sich entsprechend als Anteilswerte.

Semantische Klasse	**Absolute Frequenz KRCs Gold-standard (n = 296)**	**Relative Frequenz Gold-standard**	**Absolute Frequenz Extrak-tions-korpus (n = 340)**	**Relative Frequenz Extrak-tions-korpora**	**Prüf-größe**	**p**	**KI (95%)**
Allgemein	**131**	**0,4426**	**113**	**0,3324**	**7,6695**	**0,0056**	**[0,0315, 0,1889]**
Kommuni-kation	**41**	**0,1385**	**16**	**0,0471**	**15,1198**	**0,0001**	**[0,0430, 0,1399]**
Gesellschaft	*32*	*0,1081*	*42*	*0,1235*	*0,2314*	*0,6305*	*[-0,0683, 0,0375]*
Veränderung	*21*	*0,0709*	*40*	*0,1176*	*3,4597*	*0,0629*	*[-0,0949, 0,0015]*
Kognition	*20*	*0,0676*	*26*	*0,0765*	*0,0778*	*0,7803*	*[-0,0523, 0,0345]*
Lokation	*17*	*0,0574*	*31*	*0,0912*	*2,1213*	*0,1453*	*[-0,0774, 0,0099]*
Schöpfung	*13*	*0,0439*	*17*	*0,0500*	*0,0300*	*0,8624*	*[-0,0421, 0,0230]*
Besitz	*11*	*0,0372*	*21*	*0,0618*	*1,5227*	*0,2172*	*[-0,0612, 0,0120]*
Kontakt	*7*	*0,0236*	*15*	*0,0441*	*1,4197*	*0,2335*	*[-0,0515, 0,0105]*
Konkurrenz$_F$	**2**	**0,0068**	**2**	**0,0059**	**-**	**1**	**[0,0829, 15,9437]**
Perzeption$_F$	*1*	*0,0034*	*4*	*0,0118*	*-*	*0,3790*	*[0,0058, 2,9031]*
Verbrauch$_F$	0	0	9	0,0265	-	0,0044	[0, 0,5731]
Gefühl$_F$	0	0	4	0,0118	-	0,1274	[0, 1,7346]

Tabelle 46: Semantische Klassen einfacher Prädikate in den KRCs des Goldstandards sowie Nicht-KRCs für das Sub-EKD

Die Tabelle zeigt, dass die im Goldstandard als potentielle Bestandteile von *knowledge patterns* annotierten Verben im Vergleich zu den Prädikaten aus den zufällig ausgewählten Sätzen aus dem Sub-EKD häufiger Verben „allgemeiner“ Natur oder aber Verben der Kommunikation sind, d. h. nur für diese beiden Merkmale können zwischen den untersuchten Datensätzen signifikante Verteilungsunterschiede festgestellt werden. Zur Gruppe der „allgemeinen“ Verben gehören etwa die Verben „sein“, „bestehen“ und „darstellen“, wohingegen „bezeichnen“, „nennen“ und „beschreiben“ Verben der Kommunikation sind. Die Tabelle zeigt auch einen minimalen Unterschied in Bezug auf Verben des Verbrauchs wie beispielsweise „benötigen“, „bedürfen“ und „versorgen“, die in den Zufallssätzen aus dem Sub-EKD jedoch auch nur insgesamt neunmal auftreten. Allgemeine Verben und solche der Kommunikation machen im Sub-EKD über die Hälfte der Bausteine für *knowledge patterns* aus, für die semantische Klassen gefunden wurden, wohingegen diese beiden Klassen lediglich 40% der Prädikate in den Vergleichsdaten aus den Extraktionskorpora ausmachen.

Für den numerischen Vergleich der Muster-Verben in den beiden Datensätzen soll zunächst verglichen werden, inwieweit zwischen diesen Überlappungen existieren. Dafür werden alle Fundstellen (insgesamt 310, da nicht-prädikative Elemente sowie komplexere Prädikate wiederum nicht berücksichtigt werden können) einfacher Prädikate für die KRCs des Sub-EKD herangezogen. Mit Hilfe eines einfachen Perl-Skripts kann eine Liste der Häufigkeiten dieser Prädikate mit der Häufigkeitsliste für die generischen Sätze des Extraktionskorpus‘ verglichen werden (362 Verb-Okkurrenzen). Der Vergleich ergibt, dass in beiden Datensätzen insgesamt 252 unterschiedliche Verbtypen (Lemmata) als Prädikate bzw. Bausteine für *knowledge patterns* identifiziert wurden. Von diesen 252 Lemmata erscheinen jedoch 150 allein in der Stichprobe aus den Extraktionskorpora, was darauf hindeutet, dass diese eine wesentlich größere Variationsbreite als der untersuchte Teil des Goldstandards aufweist. Für diesen sind lediglich 53 Verben spezifisch, d. h. sie treten nur in diesem auf. Im Ergebnis teilen die beiden Datensätze lediglich 49 Verbtypen (Lemmata), die allerdings 198 (64%) der im Goldstandard annotierten und 155 (43%) der in den Referenzsätzen annotierten Fundstellen einfacher Prädikate ausmachen. Weitere quantitative Daten können mit Hilfe von zipfR (Evert/Baroni 2007) erhoben werden und sind in Tabelle 47 zusammengefasst.

Merkmal	**Goldstandard (n = 310)**	**Extraktionskorpus (n = 362)**
Anzahl unterschiedlicher Verbtypen (Lemmata) pro Datensatz	102	199
Durchschnittliche Häufigkeit jedes Typs (Lemmas)	3,0392	1,8191
Anzahl von Verbtypen (Lemmata), die nur einmal auftreten	57	142
Anteil von Verbtypen (Lemmata), die nur einmal auftreten (bezogen auf die Gesamtzahl *n* der Verb-Okkurrenzen)	0,1839	0,3923

Tabelle 47: Quantitative Typologie einfacher Prädikate in KRCs und Nicht-KRCs aus dem Sub-EKD

Vorstehende Ausführungen und die Tabelle zeigen, dass der deutsche Goldstandard auf eine relativ kleine Gruppe von 102 Verben konzentriert ist, von denen jedes im Durchschnitt dreimal auftritt. Die Stichprobe aus den Extraktionskorpora enthält dagegen fast doppelt so viele Verbtypen, von denen die deutliche Mehrheit nur ein einziges Mal in den untersuchten Sätzen erscheint (*hapax legomenon*, Bußmann 2002[3]: 302). Entsprechend ist der Anteil von *hapax legomena* in den Vergleichssätzen mehr als doppelt so hoch wie im Goldstandard und macht fast 40% der Prädikate in den Vergleichssätzen aus. Andererseits wird auch deutlich, dass sich die miteinander verglichenen Datensätze – im Hinblick auf die Anzahl der verwendeten Verbtypen – zwar deutlich voneinander unterscheiden, 64% der im Goldstandard identifizierten einfachen Prädikate jedoch nicht allein in diesem vorkommen. Vor diesem Hintergrund erscheint die musterbasierte Extraktion wissenshaltiger Kontexte als zweifelhaftes Unterfangen.

Tabelle 48 fasst die Ergebnisse der semantischen Analyse für das Russische zusammen. Aufgrund des großen Umfangs der Daten werden nur diejenigen semantischen Klassen aufgeführt, für die signifikante Verteilungsunterschiede dargestellt werden können, so dass die in der Tabelle

zusammengefassten Einzelwerte in der Summe von der Gesamtzahl *n* der pro Korpus untersuchten Verben abweichen.

Semantische Klasse (Kurzbezeichnung)	**Deutsche Umschreibung**	**Absolute Frequenz KRCs Goldstandard (n = 230)**	**Absolute Frequenz Extraktionskorpus (n = 199)**	**Prüfgröße**	**p**	**KI (95%)**
глаголы со значением связанности	**Relationsverben**	**91**	**27**	**34,8702**	**0**	**[0,1762, 0,3438]**
сохранение, пользование	**Erhalt, Verwendung**	**26**	**7**	**8,0469**	**0,0046**	**[0,0249, 0,1308]**
воля	**Wille**	**18**	**2**	**9,6864**	**0,0019**	**[0,0261, 0,1103]**
выбор, проверка, проба, оценка, измерение	*Auswahl, Überprüfung, Test, Bewertung, Messung*	*3*	*18*	*12,1198*	*0,0005*	*[-0,1246, -0,0303]*
связки и полузнаменательные глаголы	*Kopula und semantisch leere Verben*$_F$	*1*	*9*	*-*	*0,0071*	*[0,0021, 0,6784]*
информация, передача, получение сообщения	*Information, Übermittlung, Erhalt einer Nachricht*$_F$	*0*	*8*	*-*	*0,0020*	*[0, 0,4951]*

Tabelle 48: Semantische Klassen einfacher Prädikate in den KRCs des Goldstandards sowie Nicht-KRCs für das Sub-EKR

Tabelle 48 macht deutlich, dass der Goldstand im Vergleich zu den untersuchten Referenzdaten signifikant mehr Relationsverben (z. B. *состоять*, „bestehen (aus)“, *служить*, „dienen (für)“) sowie Verben, welche die Verwendung einer Sache beschreiben (z. B. *использовать*, „benutzen“), und solche Verben, die bewusste kognitive Prozesse abbilden (allein das Verb *предназначить*, „vorsehen“), enthält, jedoch in Bezug auf die Klassen Überprüfung/Messung (z. B. *контролировать*, „überprüfen“, *оценивать*, „bewerten“), Kopula (z. B. *являться*, „sein“) sowie Information (z. B. *попросить*, „bitten“, *отметить*, „anmerken“) hinter dem Referenzmaterial zurückbleibt.

Die für den Goldstandard eher typischen Verben gehören teilweise zu den bereits identifizierten Bausteinen für *knowledge patterns* bzw. Boostwörtern. Generell scheinen die für den Goldstandard häufigen semantischen Klassen die funktionalen Kriterien, die der Goldstandard-Annotation zugrunde lagen, widerzuspiegeln. Den für die Referenzdaten eher typischen semantischen Klassen gehören dagegen überwiegend Prozess- und Handlungsverben sowie Kopula an, die bislang nicht mit wissenshaltigen Kontexten in Verbindung gebracht oder als Stopwörter identifiziert wurden.

Der Vergleich der konkreten Lemmata der Musterkandidaten in den beiden Datensätzen ergibt, dass insgesamt 239 Lemmata auftreten (319 Verb-Okkurrenzen in den KRCs und 284 Verb-Okkurenzen in

den Nicht-KRCs aus dem Sub-EKR), von denen 55 nur im Goldstandard und 141 nur in den Referenzdaten erscheinen. Lediglich 43 Verben treten in beiden Datensätzen auf, machen jedoch 214 Fundstellen im Goldstandard (67%) und 93 Fundstellen (33%) in den Referenzdaten aus. Ähnlich wie im Fall des Deutschen zeigt sich hiermit eine Tendenz zu einer Einschränkung bezüglich der verwendeten Verbtypen für den Goldstandard und einer erheblich höheren Variationsbreite in den Referenzdaten, die auch durch untenstehende Tabelle 49 untermauert wird.

Merkmal	**Goldstandard (n = 319)**	**Extraktionskorpus (n = 284)**
Anzahl unterschiedlicher Verbtypen (Lemmata) pro Datensatz	98	184
Durchschnittliche Häufigkeit jedes Typs (Lemmas)	3,2551	1,5435
Anzahl von Verbtypen (Lemmata), die nur einmal auftreten	56	133
Anteil von Verbtypen (Lemmata), die nur einmal auftreten (bezogen auf die Gesamtzahl *n* der Verb-Okkurrenzen)	0,1755	0,4683

Tabelle 49: Quantitative Typologie einfacher Prädikate in KRCs und Nicht-KRCs aus dem Sub-EKR

Allerdings machen die referierten Daten auch deutlich, dass das Gros der in den KRCs identifizierten Verben nicht allein in diesen verwendet wird.

5.2.2.6 Zwischenfazit

Die Identifikation und Analyse von Bausteinen für *knowledge patterns* bestätigte zunächst für beide Sprachen eine Reihe von Verben, die bereits bei der quantitativen Auswertung von Lemma-Frequenzen als Boostwörter identifiziert worden waren. Deutlich machen obenstehende Analysen jedoch auch, dass nicht nur Verben, sondern auch syntaktische (Relativsätze, Appositionen) und typographische Elemente (Doppelpunkt, Parenthesen) sowie nicht-prädikative lexikalische Einheiten wie etwa Präpositionen die Verbindung zwischen Definiendum und Definiens herstellen und mithin Kandidaten für *knowledge patterns* sind. Zum Teil wurden auch diese Aspekte bereits in der Literatur beschrieben (z. B. verweisen auch Muresan/Klavans (2002) kurz auf Appositionen und Relativsätze, Del Gaudio/Branco (2007) nennen Interpunktionssignale und Appositionen). Allerdings scheinen die statistischen Analysen darauf hinzudeuten, dass es sich bei recht vielen der manuell identifizierten Musterkandidaten insofern um unsichere Muster handelt, als statistisch signifikante Verteilungsunterschiede zwischen dem Goldstandard und den Vergleichsdaten für eine Reihe von Lemmata nicht nachgewiesen werden konnten, das diskriminatorische Potential der genannten Elemente mithin fraglich ist. Diese Hypothese wird auch durch die Untersuchung von lexikalischen Überlappungen zwischen Goldstandard und Referenzdaten gestützt, die zeigt, dass der Goldstandard zwar tatsächlich in Teilen durch eine kleine Menge relativ typischer Verben (geringere Anzahl an Verbtypen mit durchschnittlich höherer Frequenz) beschrieben werden kann, wohingegen die Referenzdaten durch eine weitaus stärker ausgeprägte lexikalische Variation (größere Menge an Verbtypen und *hapax legomena*) geprägt sind (was die in Kapitel 3 zitierten Aussagen Barnbrooks

(2002) zur lexikalischen Uniformität von Definitionen bestätigt), die Überlappungen zwischen beiden Datensätzen jedoch beträchtlich sind und es sich bei den Kandidaten für Musterbausteine, die im Goldstandard annotiert wurden, um relativ hochfrequente Verben zu handeln scheint, die in einer Vielzahl unterschiedlicher Kontexte auftreten können. Anders formuliert: Sowohl der russische als auch der deutsche Goldstandard beschränken sich – die Einschränkungen der hier beschriebenen Methodik (manuelle Auswahl einfacher Prädikate und Abgleich mit semantischen Ressourcen mit beschränkter Abdeckung) sind hier mitzudenken – auf eine kleine Anzahl von Prädikaten (geringe Produktivität der verwendeten Verbtypen, die wohl auf das begrenzte Spektrum versprachlichter semantischer Relationen zurückzuführen ist), deren Generizität allerdings auch darin sichtbar wird, dass sie in der Mehrzahl der Fälle nicht allein im Goldstandard bzw. den Goldstandard-KRCs erscheinen (geringes diskriminatorisches Potential der *knowledge patterns* allein).
In Bezug auf die semantischen Klassen der im Goldstandard auftretenden Verben lassen die deutschen und russischen Ergebnisse keine direkten Schlüsse zu, da die den verwendeten Ressourcen zugrundeliegenden Klassifikationsschemata deutlich voneinander abweichen. Tatsächlich scheint der Goldstandard in beiden Sprachen jedoch insofern allgemeine Verben zu enthalten, als diese Beziehungen allgemeiner Natur zwischen Begriffen ausdrücken, zudem erscheint mit dem Verb „sein“ das deutsche Pendant zu dem von Drescher (1992) als allgemeines Verb *par excellence* genannten *être*[122], wohingegen zu den russischen Kandidaten für *knowledge patterns* mit *иметь* die direkte Übersetzung des von Drescher (ebd.) genannten *avoir* zählt. Allgemeinheit kann für einen großen Teil der identifizierten Konnektor-Verben in beiden Sprachen auch dahingehend konstatiert werden, dass für das Gros der identifizierten Verb-Okkurrenzen tatsächlich semantische Klassen in allgemeinsprachlichen, d. h. nicht-terminologischen Sprachressourcen gefunden werden konnten, von semantischer „Leere“ kann allerdings nicht gesprochen werden, wenn bedacht wird, dass es sich bei vielen der untersuchten Verben – sofern der russischen Klassifikation nach Švedova (2007) gefolgt wird – um relationale Verben handelt. Des Weiteren fällt ein sprachenpaarspezifischer Unterschied ins Auge, der auf dem besonderen Umgang des Russischen mit dem Verb *быть* („sein“) beruht und dazu führt, dass reine Kopula im Vergleich etwa zum deutschen „sein“ seltener erscheinen und in Teilen durch extralinguistische Elemente (z. B. einen Bindestrich) ersetzt werden.

Die Referenzdaten enthalten dagegen eher konkrete Prozess- und Handlungsverben sowie Verben der sozialen Interaktion. Neben den genannten allgemeinen bzw. Relationsverben scheinen auch solche Verben, die direkt die definierten semantischen Relationen verbalisieren, sowie – in geringerem Maße – Verben, mit deren Hilfe Eigenschaften zugeschrieben werden können, und Kommunikationsverben wie „nennen“ oder „bezeichnen“ für wissenshaltige Kontexte charakteristisch zu sein.

5.2.3 Strukturelle und grammatische Merkmale

5.2.3.1 Methode

Neben lexikalischen Merkmalen wissenshaltiger Kontexte war eingangs auch nach grammatischen und strukturellen Aspekten wie der Realisierung und Thematisierung des Definiendums, der

[122] Für das Russische ist das Auftreten von *быть* („sein“) im Tempus Präsens, wie dargestellt, aus grammatischen Gründen in der Regel nicht möglich, allerdings wurden auch für das Russische mit Kopula wie z. B. *представлять собой* sowie dem Bindestrich Verbindungselemente bzw. Kandidaten für *knowledge patterns* identifiziert, denen Generizität zugeschrieben werden kann.

Realisierung des Definitionsprädikats, ungewöhnlichen syntaktischen Formen (z. B. unpersönliche Sätze und Präsentative) sowie phraseologischen Gesichtspunkten (Geltungsrestriktionen) und der Länge und syntaktischen Komplexität wissenshaltiger Kontexte gefragt worden. Diese Aspekte sollen nun eingehender untersucht werden.

Eigenschaften der Verben können für das Deutsche dem in Kapitel 4 beschriebenen XML-Annotationsformat entnommen werden: Mittels einer einfachen XPath-Abfrage[123] können aus den annotierten Daten die morphologischen Informationen der als Verben annotierten *tokens* extrahiert werden. Diese können wiederum zu Vergleichsdaten aus den als Referenz herangezogenen *web*-Korpora sowie den zufällig ausgewählten Sätzen aus den Extraktionskorpora in Beziehung gesetzt und mit Hilfe der bereits beschriebenen und mehrfach angewandten statistischen Verfahren ausgewertet werden. Für das Russische können dagegen morphologische Informationen direkt aus der TreeTagger-Annotation gewonnen werden. Dieser globale Vergleich der morphologischen Merkmale wird wiederum für alle 632 deutschen bzw. 560 russischen KRCs gemeinsam unter Bezugnahme auf entsprechend große Zufallsstichproben von Nicht-KRCs aus den genannten Korpora durchgeführt. Allerdings ergibt dieses Vorgehen keine Informationen zur grammatischen Realisierung speziell der Definitionsprädikate (sondern vielmehr aller als Verben annotierten *tokens* in den untersuchten Sätzen), so dass diese für die Goldstandards der Entwicklungskorpora Sub-EKD und Sub-EKR manuell ausgewertet werden sollen, wobei – soweit erforderlich – auch syntaktische Aspekte berücksichtigt werden.

Die Realisierung des Definiendums wird mittels einer manuellen Analyse der bei der Erstellung des Goldstandards identifizierten Zieltermini (vgl. Kapitel 4) für das Sub-EKR und das Sub-EKD untersucht. Konkret werden dabei sowohl die grammatisch-syntaktische Realisierung des Definiendums (Bestimmung der Satzgliedeigenschaften) beschrieben als auch Thematisierungsstrategien beobachtet. Hierbei ist zu berücksichtigen, dass ein derart komplexer Vorgang wie die Thematisierung von Satzbestandteilen, bei dem – wenn er vollständig untersucht werden soll – auch kontextuelle und phonetische Aspekte, die aufgrund der hier gesammelten Daten gar nicht analysiert werden können, zu beachten sind, in seiner Gänze im Rahmen vorliegender Arbeit kaum zu untersuchen ist. Stattdessen werden lediglich die Satzanfänge in den genannten Teilen des Goldstandards betrachtet.

Die syntaktische Komplexität von Sätzen ist ebenfalls eine mehrdimensionale, komplexe Größe, an die im Rahmen vorliegender Untersuchung lediglich eine Annäherung stattfinden kann. Dafür werden die 632 KRCs des deutschen Goldstandards mit Hilfe des Stanford-Parsers (Rafferty/Manning 2008) syntaktisch annotiert. Im Unterschied zu dem in Kapitel 4 verwandten mate-*parser* (Bohnet 2010) gibt dieser *parser* Syntaxbäume aus. Ein Beispiel für einen solchen Baum ist in Abbildung 19 dargestellt.

Aus dieser Darstellung kann mit Hilfe eines Perl-Skriptes für jeden Satz die maximale Einbettungstiefe der Konstituenten ausgelesen werden, d. h. für jeden Satz kann bestimmt werden, wie tief die am tiefsten eingebettete Konstituente im Satz in den Syntaxbaum eingebettet ist. Diese Werte können wiederum zu Vergleichsdaten in Beziehung gesetzt und mittels statistischer Tests mit diesen verglichen werden. Freilich handelt es sich bei einer solchen Analyse lediglich um die Untersuchung

[123] Hierfür kann z. B. das Perl-Modul XML::XPath verwendet werden: http://search.cpan.org/~msergeant/XML-XPath-1.13/XPath.pm , Zugriff am 07.04.2014. Die entsprechende Abfrage lautet: $xp->find('corpus/sentence/token[substring(@pos, 1, 1)="V"]/@morph') und gibt das Morphologie-Attribut derjenigen *token* zurück, deren POS-*tag* mit einem V beginnt, also die morphologischen *tags* aller Verbformen.

```
(ROOT
 (S
  (PP (APPR Als) (NN Offshorebauwerk))
  (VAFIN werden)
  (VP
   (NP-OA (ADJA feststehende) (NN Bauwerke))
   (VVPP bezeichnet) ($, ,)
   (S (PRELS-SB die)
    (VP
     (PP (APPR in) (ART der) (ADJA offenen) (NN See))
     (PP (APPR vor) (ART der) (NN Küste))
     (VVPP errichtet))
    (VAFIN wurden)
    (VP (ADV -LRB-) (ADV beispielsweise)
     (CNP-OA (NN Bohrinseln) (KON und) (NN Windenergieanlagen))
     (VVPP -RRB-))))
  ($. .)))
```

Abbildung 19: Ausgabe des Stanford-Parsers

eines isolierten Anhaltspunktes für die syntaktische Komplexität der untersuchten Sätze, die hier, wie bereits festgestellt, aus Platzgründen nicht vollständig durchgeführt werden kann. Insofern als die Untersuchung der maximalen Einbettungstiefe dennoch Hinweise auf die syntaktische Struktur der untersuchten Sätze gibt, scheint sie ungeachtet dieser reduzierten Konzeptualisierung der syntaktischen Komplexität trotzdem als sinnvoll. Da allerdings nach Kenntnis der Autorin ein ähnlicher *parser* für das Russische nicht verfügbar ist, wird die Analyse nur für die deutschen Daten durchgeführt.

Die Länge der KRCs lässt sich ebenfalls mit Hilfe eines einfachen Perl-Skripts auslesen, wobei nicht-alphabetische und nicht-numerische Elemente nicht beachtet und durch Bindestrich miteinander verbundene Einheiten getrennt gezählt werden. Auch auf diese Daten können wiederum statistische Tests angewandt werden. Für die statistische Auswertung wird wieder die Statistik-Software R (R Development Core Team 2012) verwendet.

Für die Suche nach phraseologischen Elementen werden insbesondere Präpositionalphrasen ausgewertet, da im Rahmen der Untersuchung von Definitionsprädikaten und Definienda Nominal- und Verbalphrasen bereits intensiv untersucht werden. Dafür wird mit Hilfe eines Perl-Skriptes ein sehr simpler Algorithmus implementiert, der Präpositionalphrasen aus wortartenannotierten Daten ausliest. Abbildung 20 stellt den Algorithmus in verallgemeinerter Form (Pseudocode) dar.

```
foreach Zeile z in Korpus k do:
        if z enthält eine Präposition then
                output .= z;        # füge z zur Outputvariable output hinzu
                z := z+1;           # setze einen Zeiger auf die Zeile nach z
                while z enthält kein Nomen do:          # die Phrase endet mit einem Nomen
                        output .= z;
                end while;
                z := z+1;
        end if;
        if z enthält ein Nomen then
                output .= z;
        end if;
        gib output aus;
end.
```

Abbildung 20: Algorithmus für die Extraktion von Präpositionalphrasen

Freilich werden auf diese Weise nicht nur potentielle Geltungsrestriktionen, sondern alle in den Daten vorhandenen Präpositionalphrasen extrahiert. Eine manuelle Analyse, bei der terminologische Einheiten wiederum nicht berücksichtigt werden, soll alsdann diejenigen Einheiten herausfiltern, die für die kontrastive Beschreibung wissenshaltiger Kontexte möglicherweise von Interesse sein könnten. Zudem wird ein quantitatives Kriterium angesetzt, d. h. die phraseologischen Einheiten müssen eine bestimmte Mindestfrequenz aufweisen. Aufgrund der geringen Größe der untersuchten Datensätze wird hierfür der Wert 2 gewählt. Zusätzlich zu der automatischen Extraktion von Phrasen werden auch klassische Korpuskonkordanzen genutzt, um Informationen über interessierende Phrasen zu extrahieren. Auch diese Analyse wird für beide Sprachen wiederum für die KRCs der Entwicklungs- und Testkorpora gemeinsam im Vergleich zu entsprechend großen Referenzdatensätzen durchgeführt.

5.2.3.2 Verbformenanalyse

5.2.3.2.1 Russisch

Tabelle 50 fasst die Ergebnisse der morphologischen Analyse für alle Verben des russischen Goldstandards im Vergleich zu den zwei Referenzdatensätzen in der bereits mehrfach angewandten Art und Weise zusammen. Die Tabelle zeigt statistisch signifikante und überwiegend auch relevante Effektstärken (die meisten Konfidenzintervalle beziffern den minimalen Unterschied zwischen den Anteilen der jeweiligen Merkmale mit mehr als 5%) aufweisende Verteilungsunterschiede der untersuchten Merkmale zwischen dem Goldstandard und der RIC-Stichprobe, die auch überwiegend für den Vergleich des Goldstandards mit der Zufallsstichprobe aus den Extraktionskorpora Sub-EKR und Sub-TKR bestätigt werden können. Die Unterschiede für die Merkmale „perfektiver Aspekt“ (Konfidenzintervall KI = [-0,4981, -0,4208]), „imperfektiver Aspekt“ (KI = [0,4279, 0,5050]), „Imperativ“ (KI = [-0,0577, -0,0265]), „Partizip“ (Konfidenzintervall KI = [-0,1781, -0,1169]), „2. Person“ (KI = [-0,0615, -0,0288]) und „Singular“ (KI = [0,0253, 0,1123], die Konfidenzintervalle sind in der Tabelle nicht dargestellt, nur die Ergebnisse des Hypothesentests werden angegeben) sind

Tag	Absolute Frequenz Gold-standard (n = 1098)	Absolute Fre-quenz RIC (n = 1283)	Relative Frequenz Gold-standard	Rela-tive Fre-quenz RIC	Prüfgröße	p	KI (95%)
Aspekt							
*perfektiv**	*238*	*808*	*0,2168*	*0,6298*	*408,0662 (450,7237)*	*< 0,0001 (< 0,0001)*	*[-0,4498, -0,3762]*
imperfektiv*	**858**	**472**	**0,7814**	**0,3679**	**408,6745 (463,4926)**	**< 0,0001 (< 0,0001)**	**[0,3767, 0,4503]**
Modus							
Indikativ	*611*	*772*	*0,5565*	*0,6017*	*4,7915*	*0,0286*	*[-0,0859, -0,0046]*
*Imperativ**	*10*	*25*	*0,0091*	*0,0195*	*3,7124 (31,3522)*	*0,0540 (< 0,0001)*	*[-0,0206, -0,0001]*
Aktiv	*632*	*897*	*0,5755*	*0,6991*	*38,7677*	*< 0,0001*	*[-0,1629, -0,0842]*
Passiv*	**238**	**127**	**0,2168**	**0,0990**	**62,3199 (14,4155)**	**< 0,0001 (0,0001)**	**[0,0876, 0,1480]**
Gramma-tische Form							
*Infinitiv**	*82*	*244*	*0,0747*	*0,1902*	*65,8157 (32,0648)*	*< 0,0001 (< 0,0001)*	*[-0,1429, -0,0881]*
*Partizip**	*79*	*193*	*0,0719*	*0,1504*	*35,2414 (92,6093)*	*< 0,0001 (< 0,0001)*	*[-0,1041, -0,0528]*
Gerundium (Adverbial-partizipien)	*17*	*49*	*0,0155*	*0,0382*	*10,4942*	*0,0012*	*[-0,0363, -0,0091]*
*1. Person**	*1*	*89*	*0,0009*	*0,0694*	*74,3668 (-)*	*< 0,0001 (0,0172)*	*[-0,0833, -0,0536]*
*2. Person**	*12*	*47*	*0,0109*	*0,0366*	*15,1299 (32,6898)*	*0,0001 (< 0,0001)*	*[-0,0385, -0,0129]*
3. Person*	**591**	**405**	**0,5383**	**0,3157**	**119,5548 (15,8006)**	**< 0,0001 (0,0001)**	**[0,1828, 0,2624]**
Tempus							
Präsens*	**775**	**535**	**0,7058**	**0,4170**	**198,2847 (84,5087)**	**< 0,0001 (< 0,0001)**	**[0,2499, 0,3278]**
*Präteritum**	*214*	*399*	*0,1949*	*0,3110*	*41,1045 (7,4775)*	*< 0,0001 (0,0062)*	*[-0,1514, -0,0807]*
*Futur**	*13*	*80*	*0,0118*	*0,0624*	*38,8885 (16,1110)*	*< 0,0001 (0,0001)*	*[-0,0661, -0,0350]*
Numerus							
Singular*	**604**	**653**	**0,5501**	**0,5090**	**3,8523 (9,7271)**	**0,0497 (0,0018)**	**[0,0001, 0,0821]**
Plural	**393**	**337**	**0,3579**	**0,2627**	**24,8073**	**< 0,0001**	**[0,0572, 0,1333]**

Tabelle 50: Ergebnisse der kontrastiven Analyse morphologischer Informationen für russische Verben der Goldstandard-KRCs im Vergleich zu Referenzdaten

zwischen dem Goldstandard und der Stichprobe aus den Extraktionskorpora sogar stärker ausgeprägt als zwischen Goldstandard und RIC. Diese Ergebnisse deuten darauf hin, dass z. B. Imperative für die

Extraktionskorpora selbst eher typisch sind (zumindest das Sub-EKR enthält, wie in Kapitel 4 dargestellt, einen hohen Anteil von Betriebsanleitungen, in denen Imperative als sprachliches Korrelat der Anleitung zur Durchführung bestimmter Handlungen auftreten), für den Goldstandard selbst jedoch nicht als häufig angesehen werden können. Für das Merkmal Passiv ergibt die Studie zwar für beide Vergleichsdatensätze signifikante Verteilungsunterschiede, hinsichtlich des Merkmals Aktiv kann der Unterschied für die Stichprobe aus den Extraktionskorpora Sub-EKR und Sub-TKR jedoch nicht bestätigt werden. Dieses Ergebnis kann dahingehend interpretiert werden, dass Passivierung ein für die Fachsprache zwar eher typisches sprachliches Mittel darstellt (Unterschied Goldstandard-RIC), hinsichtlich des Goldstandards aber bestenfalls von einer Tendenz zur noch häufigeren Verwendung des Passivs ausgegangen werden kann. Der geringere Anteil von Infinitiven ist dagegen möglicherweise auf den bereits diskutierten geringeren Anteil von Modalverben, die Infinitive nach sich ziehen, zurückzuführen. Überraschend wirkt der geringere Anteil von Partizipien im Goldstandard. Das Russische besitzt insgesamt vier Partizipien (vgl. Bendixen/Voigt 1999), die sowohl in attributiver als auch in prädikativer Weise verwendet werden können und alle gleichermaßen als „Partizip“ getaggt werden. Entsprechend komplex ist die Suche nach funktionalen Ursachen für den empirisch beobachtbaren Unterschied. Generell kann festgehalten werden, dass die Partizipien insbesondere bei attributiver Verwendung als sprachliche Mittel der Kompression betrachtet werden können, d. h. Sachverhalte werden statt mit einem eigenen Prädikat lediglich attributiv ausgedrückt (vgl. die Beispiele in ebd.: 120). Des Weiteren kann mit Hilfe von Adjektiven, die *de facto* Ableitungen von Partizipialformen darstellen, auf bereits Bekanntes verwiesen und damit auf textueller Ebene Kohäsion erzeugt werden (z. B. *данный автомобиль*, wörtlich: „das gegebene Fahrzeug“, besser übersetzt als „dieses Fahrzeug“, *названные недостатки*, „die genannten Nachteile“, besser übersetzt als „besagte Nachteile“, *вышеописанный метод*, „die oben beschriebene Methode“). Vor dem Hintergrund dieser Betrachtungen besteht Grund zu der Annahme, dass die beschriebenen Verteilungsunterschiede im Hinblick auf Partizipien mit dem in wissenshaltigen Kontexten erwartbaren Bestreben nach größtmöglicher Explizitheit korrespondieren. Weitere Unterschiede in Bezug auf die Wahl der grammatischen Person und des Tempus‘ verweisen auf das Bestreben nach sachbezogener, gegenwärtig aktueller Information.

Abbildung 21 zeigt das Ergebnis der manuellen Wortformenanalyse für bereits identifizierte verbale Kandidaten für Musterbausteine im Goldstandard für das Sub-EKR (vgl. Abschnitt 5.2.2.5.1), d. h. für die einfachen Prädikate, mit deren Hilfe Definiendum und Definiens miteinander verbunden werden. Tabelle 51 gibt zu jedem der dargestellten Typen eine Erklärung und ein (leicht angepasstes) Beispiel aus dem Goldstandard. Die relevanten Verbformen sind jeweils durch Unterstreichung hervorgehoben.

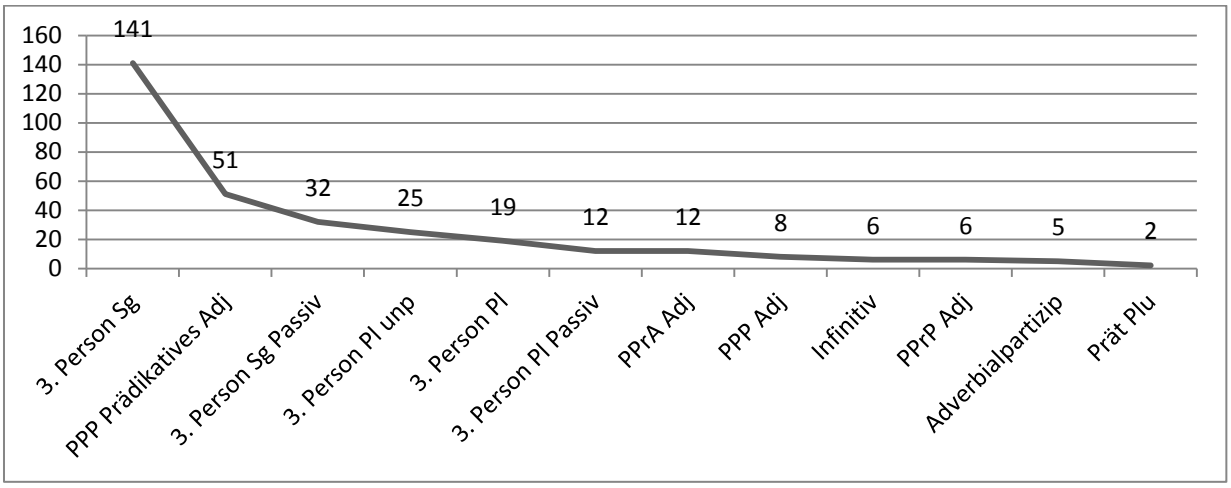


Abbildung 21: Grammatische Realisierung von *knowledge patterns* in den KRCs des Sub-EKR

Form	Erklärung	Beispiel	Übersetzung
3. Person Sg	3. Person Singular	тормозная система состоит из	die Bremsanlage besteht aus
PPP Prädikatives Adj	Partizip Präteritum Passiv als prädikatives Adjektiv	тормоз предназначен для	die Bremse ist vorgesehen für
3. Person Sg Passiv	3. Person Singular Passiv	охлаждение применяется в целях	die Kühlung wird eingesetzt, um
3. Person Pl unb	3. Person Plural, unbestimmt-persönlicher Satz	различают барабанные и дисковые тормоза	man unterscheidet Trommel- und Scheibenbremsen
3. Person Pl	3. Person Plural	в привод входят	zum Antrieb gehören
3. Person Pl Passiv	3. Person Plural Passiv	повороты руля преобразуются в	die Umdrehungen des Lenkrads werden in ... umgewandelt
PPrA Adj	Partizip Präsens Aktiv als attributives Adjektiv	тормозная система, служащая для	die Bremsanlage, die dazu dient,
PPP Adj	Partizip Präteritum Passiv als attributives Adjektiv	регуляторы давления, смонтированные на	die auf … montierten Druckregler
Infinitiv	Infinitiv	контур должен делиться	der Kreislauf muss aufgeteilt werden
PPrP Adj	Partizip Präsens Passiv als attributives Adjektiv	тормоз, приводимый в действие механически	die mechanisch betriebene Bremse
Adverbialpartizip	Adverbialpartizip	регуляторы давления корректируют давление тормозной жидкости, исключая возможность блокировки колес	die Druckregler korrigieren den Druck der Bremsflüssigkeit und verhindern so ein Blockieren der Räder

Prät Pl	Präteritum Plural	наибольшее распространение получили фрикционные тормозные механизмы	die weiteste Verbreitung fanden Friktionsbremssysteme

Tabelle 51: Beispiele für Realisierungsformen russischer *knowledge patterns*

Die manuelle Analyse bestätigt speziell für die verbalen Musterkandidaten die Ergebnisse der in Tabelle 50 präsentierten globalen Analyse: Das Gros (141 von insgesamt 319 analysierten Verben, also ca. 44%) derjenigen Verben, die als potentielle Bausteine für *knowledge patterns* markiert wurden, steht in der dritten Person Singular Präsens Indikativ und stellt damit eine direkte Beziehung zwischen dem jeweiligen Definiendum und den im KRC enthaltenen Erklärungsbestandteilen her. Werden alle vorhandenen Passivformen jedoch zusammengezählt (Labels PPP Prädikatives Adj, 3. Person Sg Passiv, 3. Person Pl Passiv, PPP Adj, PPrP Adj), zeigt sich, dass 109 von 319 und damit ca. 34% der analysierten Verben eine Passivierung ausdrücken.
Beachtung verdient des Weiteren die Kategorie der unbestimmt-persönlichen Sätze (Bendixen/Voigt 1999: 156, *неопределенно-личное предложение,* vgl. Zolotova et al. 1998: 112 ff.), auf die bereits verwiesen wurde. Generell scheint diese Art von Sätzen insofern auf einen gewissen Grad der Allgemeinheit hinzudeuten, als Informationen über den Träger der Handlung nicht gegeben werden (ebd.: 115):

> „Неопределенно-личная субъектная модификация выражает дейтсвие третьего лица, неучастника коммуникации, не названного за ненадобностью или по неведению [...].“
> [*Übersetzung:* Die unbestimmt-persönliche Subjektmodifikation beschreibt eine Handlung einer dritten Person, die nicht Teilnehmer der Kommunikation ist und nicht benannt wird, da dies nicht nötig ist oder die entsprechenden Informationen nicht bekannt sind […].“]

Bendixen und Voigt (1999: 156) verweisen darauf, dass unbestimmt-persönliche Sätze in der Regel als Passiv ins Deutsche zu übersetzen seien. Das Goldstandard-Beispiel aus Tabelle 51 macht jedoch deutlich, dass auch andere Korrespondenzen existieren, und zwar eine Korrespondenz zum deutschen Indefinitpronomen „man“ (man unterscheidet …) bzw. zu subjektlosen deutschen Konstruktionen mit nicht-phorischem „es“ (es wird unterschieden …, expletives „es“, vgl. Zifonun et al. 1997: 1082). Während Letzteres nicht als sicheres Verallgemeinerungssignal bewertet werden kann, sondern als „reine Stellungseinheit“ (ebd.: 1605) betrachtet wird, weist Ersteres durchaus auf die Allgemeingültigkeit der Aussage hin (ebd.: 939, Hervorhebungen im Original): „*Man* […] signalisiert eine Verallgemeinerung von Meinungen, Intentionen oder Urteilen.“ Eine solche verallgemeinernde Bedeutung kann den unbestimmt-persönlichen Sätzen des Russischen im hier diskutierten Kontext der Vermittlung begrifflichen Wissens durchaus auch zugeschrieben werden, wenn argumentiert wird, dass die Benennung des Handlungsträgers insofern nicht erforderlich ist, als es sich um einen kollektiven Handlungsträger – die Mehrheit der infrage kommenden Subjekte – handelt. Darüber hinaus machen insbesondere die in der Tabelle dargestellten Partizipialkonstruktionen sowie die Konstruktionen mit Adverbialpartizipien deutlich, dass Verbformen, die in russischen KRCs Begriffserklärungen einleiten, in komplexe syntaktische Konstruktionen eingebettet sein können, die in der relevanten Literatur nicht immer ausreichend berücksichtigt werden.

5.2.3.2.2 Deutsch

Tabelle 52 präsentiert die Ergebnisse der morphologischen Analyse für die deutschen Daten.

Tag	Absolute Frequenz Gold-standard (n=1003)	Absolute Fre-quenz NCL (n = 923)	Relative Frequenz Gold-standard	Relative Fre-quenz NCL	Prüf-größe	p	KI (95%)
Modus							
Imperativ$_F$	**5**	**0**	**0,0050**	**0**	**-**	**0,0633**	**[0,8448, ∞)**
Indikativ*	**989**	**847**	**0,9860**	**0,9177**	**48,9344 (11,9992)**	**< 0,0001 (0,0005)**	**[0,0482, 0,0886]**
*Konjunktiv**	*4*	*76*	*0,0040*	*0,0823*	*72,1656 (23,4506)*	*< 0,0001 (< 0,0001)*	*[-0,0975, -0,0592]*
Person							
*1. Person**	*1*	*62*	*0,0010*	*0,0672*	*64,4518 (11,4995)*	*< 0,0001 (0,0007)*	*[-0,0835, -0,0489]*
2. Person	*5*	*7*	*0,0050*	*0,0076*	*0,1886*	*0,6641*	*[-0,0107, 0,0055]*
3. Person	**992**	**854**	**0,9890**	**0,9252**	**47,5390**	**< 0,0001**	**[0,0446, 0,0830]**
Tempus							
Präsens*	**963**	**655**	**0,9601**	**0,7096**	**222,6134 (83,3541)**	**< 0,0001 (< 0,0001)**	**[0,2177, 0,2832]**
*Vergangen-heit**	*30*	*268*	*0,0299*	*0,2904*	*247,3158 (96,3116)*	*< 0,0001 (< 0,0001)*	*[-0,2926, -0,2283]*
Numerus							
Singular	*687*	*675*	*0,6849*	*0,7313*	*4,7686*	*0,0290*	*[-0,0880, -0,0048]*
Plural	**316**	**248**	**0,3150**	**0,2687**	**4,7686**	**0,0290**	**[-0,0880, -0,0048]**

Tabelle 52: Ergebnisse der kontrastiven Analyse morphologischer Informationen für deutsche Verben der Goldstandard-KRCs im Vergleich zu Referenzdaten

Ebenso wie die russischen Daten zeigen die deutschen Daten signifikante Unterschiede mit relevanten Effektstärken für die Merkmale Modus (Überwiegen des Indikativs im Goldstandard), Person (geringerer Anteil der 1. Person im Goldstandard) und Tempus (Überwiegen des Tempus Präsens im Goldstandard). Anders als im Falle des Russischen, das den Konjunktiv analytisch mit Hilfe der Partikel *бы* bildet, können mit Hilfe der beschriebenen einfachen Analysemethode auch Konjunktivbildungen erfasst und so gezeigt werden, dass auch der Konjunktiv – erwartbar – für den Goldstandard eher untypisch zu sein scheint.

Abbildung 22 zeigt nun wiederum das Ergebnis der manuellen Analyse für bereits identifizierte verbale Kandidaten für Musterbausteine im Goldstandard des Sub-EKD, entsprechende Funktionsverbgefüge und komplexere Prädikate[124]. Tabelle 53 gibt zu jedem der dargestellten Typen

[124] Funktionsverbgefüge wurden für die Analyse semantischer Klassen in Abschnitt 5.2.2.5.3 aufgrund der höheren Analysekomplexität nicht berücksichtigt, so dass die Gesamtzahl der hier untersuchten Verben von der dort angegebenen Gesamtzahl leicht abweicht.

eine Erklärung und ein Beispiel. Die relevanten Verbformen sind wieder durch Unterstreichung hervorgehoben.

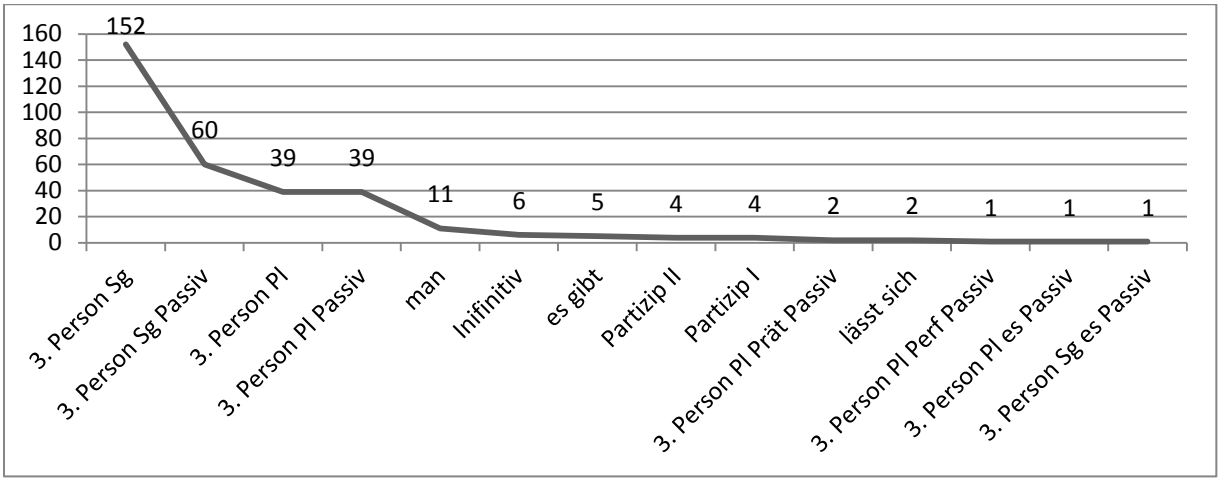


Abbildung 22: Grammatische Realisierung von *knowledge patterns* in den KRCs des Sub-EKD

Form	**Erklärung**	**Beispiel**
3. Person Sg	3. Person Singular	Windenergie hat ihren Ursprung in
3. Person Sg Passiv	3. Person Singular Passiv	das Potenzial der energetischen Nutzung wird unterschieden in
3. Person Pl	3. Person Plural	Windströmungen enthalten Bewegungsenergie
3. Person Pl Passiv	3. Person Plural Passiv	die Schwingungskräfte werden durch die hinteren Zugstränge abgefangen
man	3. Person Singular mit Indefinitpronomen „man“	grundsätzlich unterscheidet man zwischen zwei Generatortypen
Infinitiv	Infinitiv	Offshore-Fundamente können aus Beton oder Mehrbeinstrukturen bestehen
es gibt	thetische Präsentativkonstruktion mit der Wendung „es gibt“	es gibt zwei Bauformen von Windkraftanlagen
Partizip II	Partizip II	Asynchrongeneratoren, auch Induktionsgeneratoren genannt,
Partizip I	Partizip I	Rotor (bestehend aus Rotorblättern und Nabe)
3. Person Pl Prät Passiv	3. Person Plural Präteritum Passiv	wurden sogenannte Vorrangflächen für die Windenergie aufgestellt
lässt sich	Infinitivkonstruktion mit dem Hilfsverb „(sich) lassen”	Energie lässt sich definieren als
3. Person Pl Perf Passiv	3. Person Plural Perfekt Passiv	Asynchrongeneratoren sind ursprünglich als Elektromotoren konzipiert worden
3. Person Pl es Passiv	Passivsatz im Plural mit nicht-phorischem „es”	es werden drei Arten von Generatoren zur Stromerzeugung verwendet

3. Person Sg es Passiv	Passivsatz im Singular mit nicht-phorischem „es"	es wird unterschieden zwischen dem Darrieus-Typ und den Vertikalachsern

Tabelle 53: Beispiele für Realisierungsformen deutscher *knowledge patterns*

Abbildung und Tabelle zeigen, dass das Deutsche ebenso wie das Russische eine Reihe unpersönlicher Konstruktionstypen verwendet, die auch dem Ausdruck von Allgemeinheit dienen können. Der allgemeine Charakter der mit Hilfe des Indefinitpronomens „man" formulierten Aussage wird dabei im oben dargestellten Beispiel noch durch das Adverb „grundsätzlich" unterstrichen. Die Konstruktionen mit nicht-phorischem „es" dagegen dienen möglicherweise der Fokussierung auf die einzuführende kategorische Unterscheidung (Zifonun et al. 1997: 1606, Hervorhebung im Original): „Die Konstruktion mit *es* wird […] gewählt, wenn die gesamte Information zusammenhängend fokussiert werden soll." Erwähnung verdienen darüber hinaus die thetischen Präsentative auf „es gibt", die ein Mittel zur Einführung kategorischer Unterscheidungen darzustellen scheinen. In einigen wenigen Sätzen tritt zudem das Hilfsverb „lässt sich" auf, das den Musterkandidaten im Infinitiv nach sich zieht. Ähnlich wie bereits im Falle des Russischen ergibt sich auch für die deutschen Daten bei Addition aller Passivverben ein Passivanteil von ca. 32% (104 von insgesamt 327 untersuchten Verben). Allerdings steht auch hier wie im Russischen die 3. Person Singular Präsens Indikativ als Realisierungsform für die verbalen Musterkandidaten mit 152 von 327 Verben (ca. 46%) an der Spitze der Liste.

5.2.3.2.3 Zwischenfazit

Obige Ausführungen zur Verbmorphologie zeigen, dass deutsche ebenso wie russische wissenshaltige Kontexte zu nahezu 50% die dritte Person Singular des Konnektorverbs einsetzen, um ein Definiendum mit Erklärungsbestandteilen zu verbinden. Allerdings erscheinen in den untersuchten KRCs auch zahlreiche andere Verbformen, mit deren Hilfe z. T. komplexe Satzstrukturen realisiert werden. Sowohl russische als auch deutsche KRCs sind zudem weit überwiegend im Tempus Präsens formuliert und erfüllen damit ein von Drescher (1992) genanntes Generalisierungskriterium.
Darüber hinaus wurden häufige Verwendungen von Passivformen – zwar dürfte es sich hierbei auch um eine Besonderheit fachlicher Texte handeln, allerdings zeigt die Analyse der russischen Daten (für das Deutsche können Passivkonstruktionen aufgrund der analytischen Passivbildung nicht auf morphologischer Ebene beschrieben werden) auch für den Vergleich des Goldstandards mit den zufällig aus den Extraktionskorpora ausgewählten Sätzen Unterschiede im Hinblick auf die Verwendung des Passivs – und unpersönlichen Sätzen (z. B. im Russischen die unbestimmt-persönlichen Sätze oder aber im Deutschen Sätze mit indefinitem „man") festgestellt. Des Weiteren finden sich zumindest für das Deutsche in Form einiger Sätze mit nicht-phorischem „es" Präsentative. Drescher (1992) führt diese Elemente zwar als mögliche linguistische Hinweise auf allgemeine Aussagen an, untersucht sie aber selbst nicht. Im Vergleich zu den zufällig ausgewählten 337 generischen Sätzen aus dem Sub-EKD, für die ebenfalls linguistische Prädikate analysiert wurden (vgl. Abschnitt 5.2.2.5.3) zeigt sich jedoch, dass die dort identifizierten Prädikate in acht Fällen – also nur unwesentlich seltener als in den KRCs – in eine unpersönliche Konstruktion mit „man" eingebettet sind (für das Auftreten von „man" in Hauptsätzen). Entsprechend kann kaum von einer Spezifik wissenshaltiger Kontexte ausgegangen werden. Die thetischen Präsentativkonstruktionen treten in derselben Zufallsstichprobe dreimal auf, so dass auch hier nicht von einer Spezifik ausgegangen werden kann. Die unbestimmt-persönlichen Sätze sind dagegen in Bezug auf die dort annotierten

Prädikate im russischen Goldstandard mit 25 Fundstellen deutlich häufiger als in den zum Vergleich herangezogenen 292 Nicht-KRCs aus dem Sub-EKR (7 Fundstellen). Wird nun gefragt, ob es sich bei den beschriebenen Merkmalen um Verallgemeinerungssignale handelt, können zumindest auf Grundlage der deutschen Daten keine eindeutigen Schlüsse gezogen werden. Allerdings sind auch für statistisch abgesicherte Aussagen in Bezug auf das Russische noch eingehendere Studien erforderlich. Hinsichtlich des Modus' überwiegt in beiden Sprachen sehr deutlich der Indikativ. Wird mit Halliday (1994[2]) unter Modus auch das Merkmal Polarität (positive vs. negative Aussage) verstanden, kann zudem darauf verwiesen werden, dass im Zusammenhang mit der Untersuchung lexikalischer Merkmale wissenshaltiger Kontexte bereits auf den weitgehenden Verzicht auf Negationspartikeln verwiesen worden war.
Für das Russische konnte außerdem gezeigt werden, dass KRCs den imperfektiven Aspekt bevorzugen, womit die bereits im Zusammenhang mit der Suche nach Boostwörtern vermutete Präferenz für diesen Aspekt aus quantitativer Sicht bestätigt werden konnte. Des Weiteren wurde zwar für die russischen KRCs ein signifikant geringerer Anteil an Partizipialformen gezeigt, jedoch verdeutlichen die in der Diskussion angeführten Beispiele, dass KRCs in unterschiedlichste syntaktische Konstruktionen eingebettet sind und somit die der definitionsbezogenen Literatur entnehmbare Einschränkung auf den Hauptsatz bzw. den Argumentstatus des Definiendums in Bezug auf das Prädikat insbesondere für das Russische die Kategorie des wissenshaltigen Kontexts stark einengt.

5.2.3.3 Analyse des Definiendums

5.2.3.3.1 Russisch

Abbildung 23 stellt dar, welche grammatisch-syntaktische Funktion die in den 292 KRCs des Goldstandards für das Sub-EKR markierten Definienda annehmen.

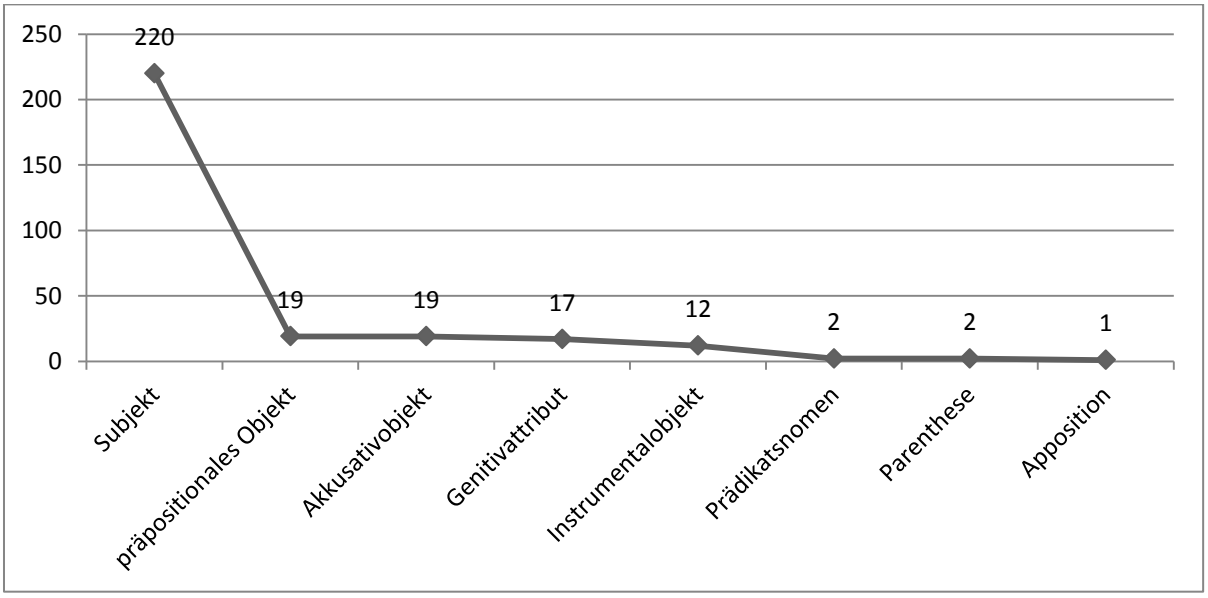


Abbildung 23: Realisierung des Definiendums in den 292 KRCs des Sub-EKR

Die Abbildung zeigt, dass 75% aller Definienda im Goldstandard des Sub-EKR als Subjekt realisiert sind. Deutlich weniger häufig sind dagegen Realisierungen als Objekt in einem obligaten Kasus

(einschließlich des Instrumentals, mit dessen Hilfe in Passivkonstruktionen die Agensangabe erfolgt) mit oder ohne Präposition oder als Attribut. Bildungen, in denen das Definiendum als Prädikatsnomen, Apposition oder innerhalb einer Parenthese auftritt, können dagegen für die hier untersuchten Daten als eher nachrangig betrachtet werden. Die nachfolgende Tabelle 54 fasst zusammen, mit Hilfe welcher syntaktischen Satzbestandteile in den untersuchten KRCs jeweils die Satzanfänge realisiert sind.

Realisierung des Satzanfangs	**Anzahl Fundstellen**
Nominalphrase	
Definiendum	201
Subjekt	32
Objekt	5
Prädikatsnomen	3
∑	*241*
Präpositionalphrase	
{с, на, в, к} + Ort	12
{к, на, в} + Definiendum	9
{в, на, при} + fachlicher Zusammenhang	9
для + Funktion	8
{в зависимости, по} + Unterscheidungskriterium	3
помимо + Angabe eines Bestandteils	1
согласно + normative Referenz	1
∑	*43*
Verbalphrase	
Prädikat	4
Modalbestimmung	2
∑	*6*
Teilsatz	
Finalsatz	1
Konditionalsatz	1
∑	*2*
	292

Tabelle 54: Realisierung der Satzanfänge in den 292 KRCs des Sub-EKR

Die Tabelle macht deutlich, dass ca. 72% (210 von 292 Sätzen) der untersuchten russischen KRCs das Definiendum durch eine satzinitiale Positionierung im Rahmen einer Nominal- oder Präpositionalphrase direkt thematisieren. Geht dem Definiendum dabei eine Präposition voraus, ist diese Argument eines Prädikats oder führt das Definiendum als erforderlichen fachlichen Zusammenhang ein[125], vgl. Beispiele 46) und 49) (Abschnitt 5.2.2.5.1). Andere Möglichkeiten der Satzeröffnung mittels einer Nominalphrase beinhalten die satzinitiale Positionierung eines Subjekts (das in diesem Fall nicht das Definiendum ist), eines Objekts oder eines Prädikatsnomens. Darüber hinaus werden mit Hilfe von Präpositionalphrasen unterschiedlichste Zusatzinformationen vermittelt, z. B. Ortsangaben (vgl. Beispiel 79), die Definienda sind in diesem und den folgenden Beispielen durch Unterstreichung hervorgehoben), fachlich-inhaltliche (häufig prozessuale) Zusammenhänge (vgl. Beispiel 80)), Funktionsangaben (vgl. Beispiel 81)), Unterscheidungsmerkmale (vgl. Beispiel 82)),

[125] Für das Deutsche werden analoge Formen in Anlehnung an die IDS-Grammatik als „Situativkomplement“ bezeichnet.

Angaben weiterer Bestandteile eines komplexen Mechanismus‘ (vgl. Beispiel 83)) oder (wenngleich in vorliegendem Beispiel unvollständige) Verweise auf existierende normative Regelungen (vgl. Beispiel 84)).

79) На концах вала расположены карданные шарниры, которые состоят из двух вилок, сочленённых крестовиной 3.
[*Übersetzung:* An den Enden der Welle sind Kreuzgelenke angebracht, die aus zwei mit Hilfe des Kreuzstücks 3 verbundenen Gelenkgabeln bestehen.]

80) При отказе одного из контуров рабочей тормозной системы используется второй контур, обеспечивающий остановку автомобиля с достаточной эффективностью.
[*Übersetzung:* Im Fall eines Defekts eines Kreises des Betriebsbremssystems wird der zweite Kreis verwendet, um das Fahrzeug mit ausreichender Effektivität zum Stehen zu bringen.]

81) Для контроля теплового состояния двигателя служат датчик температуры и показывающий прибор в кабине водителя.
[*Übersetzung:* Für die Überprüfung des Wärmezustands des Motors dienen der Temperaturgeber und die Anzeigeeinheit in der Fahrerkabine.]

82) По назначению стали делятся на конструкционные, инструментальные и стали с особыми свойствами.
[*Übersetzung:* Nach dem Verwendungszweck unterscheidet man Baustahl, Werkzeugstahl und Spezialstahl.]

83) Помимо гидравлического привода, тормозные механизмы задних колес имеют механический привод от рычага стояночного тормоза, установленного на полу кузова.
[*Übersetzung:* Neben einem hydraulischen Antrieb besitzen die Hinterradbremsen noch einen mechanischen Antrieb über den Hebel der Handbremse, der am Boden der Karosserie verankert ist.]

84) Согласно этому стандарту тормозное управление должно состоять из четырех систем: рабочей, запасной, стояночной и вспомогательной.
[*Übersetzung:* Gemäß diesem Standard muss die Bremsanlage aus vier Systemen bestehen: dem Betriebs-, dem Hilfs-, dem Hand- und dem Zusatzbremssystem.]

Die Satzeinleitung durch ein Prädikat tritt – obwohl im Russischen als einer Sprache mit freier Wortfolge grundsätzlich möglich – in den untersuchten KRCs dagegen nur selten auf (vgl. Beispiel 85)). Nicht zuletzt treten einige wenige Satzeröffnungen auf, in denen mit Hilfe einer ganzen Phrase modale Informationen an den Beginn des Satzes gestellt werden, wohingegen beispielsweise einfache Zeitangaben den jeweils übergeordneten Phrasen zugeordnet und daher nicht berücksichtigt wurden (vgl. Beispiel 23), Abschnitt 5.2.2.3.1).

85) Нам известны такие типы тормозных приводов: механический; гидравлический; пневматический; электрический; комбинированный.
[*Übersetzung:* Uns sind die folgenden Typen von Bremsantrieben bekannt: mechanisch, hydraulisch, pneumatisch, elektrisch, kombiniert.]

5.2.3.3.2 Deutsch

Wiederum analog zur Darstellung für das Russische zeigt Abbildung 24, auf welche Weise das Definiendum in den 337 KRCs des Goldstandards für das Sub-EKD realisiert ist. Auch hier zeigt sich, dass die Realisierung des Definiendums als syntaktisches Subjekt überwiegt, jedoch mit ca. 60% deutlich seltener ist als in den russischen Daten. Zu den bereits für das Russische beschriebenen Realisierungsvarianten Subjekt, Objekt, Attribut, Prädikatsnomen, Parenthese und Apposition treten hier zudem die Formen Situativkomplement, präpositionale Erweiterung, Doppelpunkt und Prädikativ hinzu. Beispiel 86) (auch hier werden die Definienda wiederum durch Unterstreichung hervorgehoben) illustriert eine Realisierung des Definiendums als Situativkomplement. Situativkomplemente decken dabei Strukturen ab, die der „[…] lokalen oder temporalen Spezifizierung von Sachverhalten bei Verben [dienen], die einen entsprechenden situativen Parameter implizieren“ (Zifonun et al. 1997: 1101).

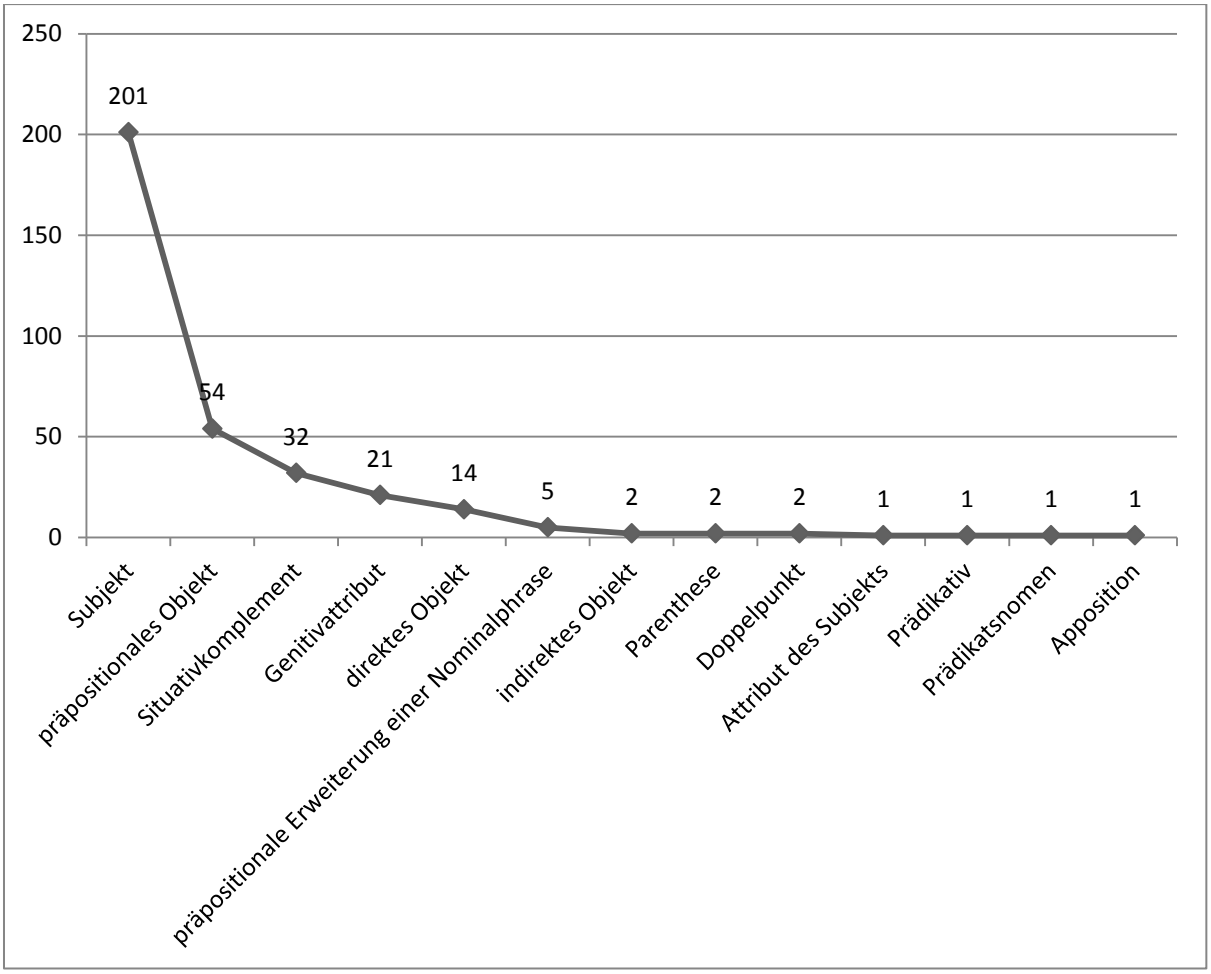


Abbildung 24: Realisierung des Definiendums in den 337 KRCs des Sub-EKD

86) Bei der Jacket-Gründung wird der Turm der WEA auf einer Fachwerkstruktur aus Hohlprofilen verankert, die ebenfalls mit Pfählen im Meeresboden verankert wird.

Insofern, als die lokale Spezifikation im obigen Beispiel bestenfalls eine mittelbare zu sein scheint und passender als inhaltliche, d. h. sachverhaltsbezogene Konkretisierung betrachtet werden kann, wäre es auch möglich, eine solche fachliche Spezifizierung aus funktionaler Sicht als Geltungsrestriktion

(ebd.: 602) zu charakterisieren. In der Tat geben Zifonun et al. (ebd.) ein Beispiel für die Formulierung von Geltungsrestriktionen mit der satzeinleitenden Präposition „bei“. Allerdings finden sich in den untersuchten KRCs auch strukturähnliche Formulierungen mit anderen Präpositionen mit dezidiert lokaler Bedeutung (vgl. Beispiel 68), Abschnitt 5.2.2.5.2). Zudem soll der gewählte Fokus auf grammatisch-syntaktische Aspekte beibehalten werden, so dass die entsprechenden Strukturen auch weiterhin als Situativkomplemente bezeichnet werden. Als präpositionale Erweiterungen von Nominalphrasen wurden solche Definienda eingestuft, die durch eine Präposition mit einer sie dominierenden Nominalphrase verbunden sind (vgl. Beispiel 87)).

87) Es gibt im Wesentlichen zwei Bauformen von Windkraftanlagen: solche mit senkrechter Drehachse und solche mit waagrechter Achse.

Die Grammatik von Zifonun et al. (ebd.: 1975 ff.) klassifiziert diese Formen als Nomenkomplemente. Grundsätzlich sind auch Realisierungen ohne oder mit schwächerer Argumentselektion, d. h. Supplemente, möglich, wobei diese feinkörnige Unterscheidung für das hier verfolgte Forschungsvorhaben jedoch nur von untergeordneter Bedeutung ist. Aus diesem Grund werden die betreffenden Strukturen auch weiterhin als präpositionale Erweiterungen von Nominalphrasen bezeichnet.
Die Verwendung von Doppelpunkten war bereits im Zusammenhang mit der Untersuchung von Kandidaten für Musterbausteine untersucht worden. Die hier beschriebene Klassifizierung bezieht sich dabei auf Beispiele, in denen das Definiendum als einziges Element der Begriffserklärung vorangestellt und von dieser durch einen Doppelpunkt getrennt ist. Ein Beispiel für eine solche Struktur findet sich in Beispiel 66) (Abschnitt 5.2.2.5.2). Als Prädikativ wurden schließlich solche Formen charakterisiert, die – im Gegensatz zu Prädikatsnomen – zwar vom Prädikat abhängig, jedoch keine Nomina sind. Ein Beispiel hierfür ist Satz 88).

88) Bei modernen Windkraftanlagen zur Stromerzeugung sind die Rotorflügel aerodynamisch ausgelegt, das heißt, sie haben ein aerodynamisches Profil, das einen Auftrieb erzeugt, der in ein Drehmoment umgesetzt wird, das zum Antrieb des Generators dient.

Tabelle 55 stellt dar, mit Hilfe welcher syntaktischen Konstruktionen die Satzanfänge in den untersuchten KRCs realisiert sind.

Realisierung des Satzanfangs	**Anzahl Fundstellen**
Nominalphrase	
Definiendum	173
Subjekt	45
Prädikatsnomen	11
Genus Proximum	8
Kardinalität der Aufzählung	1
$\sum$	*238*
Präpositionalphrase	
{bei, in, mit, unter, durch, als, mit Hilfe, über, von, zu} + Definiendum	41
{bei, in, für, hinsichtlich, von, nach, durch} + fachlicher Zusammenhang	18
{in, mittig} + Ortsangabe	5
zu + Funktion	5

in + Abgrenzung + Prädikat	2
∑	*71*
Verbalphrase	
Prädikat	12
es	9
man	1
Prädikatsadjektiv	1
∑	*23*
Teilsatz	
Konditionalsatz	3
Finalsatz	2
∑	*5*
	337

Tabelle 55: Realisierung der Satzanfänge in den 337 KRCs des Sub-EKD

Die Tabelle zeigt, dass eine satzinitiale Positionierung des Definiendums mit oder ohne Einbindung in eine Präpositionalphrase in 214 von 337 KRCs (ca. 64%) und damit wiederum seltener als in den russischen KRCs beobachtet werden konnte. Die Varianten der Satzeinleitung durch ein Prädikatsnomen oder ein g*enus proximum* weisen zwar formale Ähnlichkeiten auf, jedoch ist das satzinitiale Nomen nur im Falle der Einleitung durch ein g*enus proximum* ein echter Oberbegriff des nach rechts verschobenen Definiendums (vgl. Beispiel 89)), so dass eine invertierte Definitionsstruktur entsteht, die den von Pearson (1998, vgl. Abschnitt 2.3.2.3) beschriebenen englischen Strukturen ähnelt:

89) Das Systemelement, mit dem die im Wind enthaltene Energie in eine mechanische Drehbewegung umgewandelt wird, bezeichnet man bei einer modernen Windkraftanlage als Rotor.

In einem Fall enthält der Satzanfang auch eine direkte Angabe der Kardinalität der folgenden Aufzählung, die zur Darstellung kategorialer Unterschiede dient (vgl. Beispiel 90, sic!):

90) Drei technische Lösungen sind bei der Blattverstellung möglich: mechanisch, hydraulisch und elektrisch:

Präpositionalphrasen am Satzbeginn vermitteln, wenn sie nicht das Definiendum einführen, unterschiedlichste Zusatzinformationen mit fachlichem Bezug. Da es sich bei diesen Konstruktionstypen zumindest in dem untersuchten Datensatz um Kategorien mit nur wenigen Beispielen handelt, wurden auch Sätze, in denen auf diese Weise agensähnliche Bestandteile von Passivkonstruktionen (vgl. Beispiel 91)) und Unterscheidungsmerkmale (vgl. Beispiel 30), Abschnitt 5.2.2.3.2) benannt werden, dieser Gruppe zugeordnet.

91) Durch Energieentzug aus dem Wind wird der Rotor (bestehend aus Rotorblättern und Nabe) in Drehung versetzt.

Als für den betrachteten Datensatz eher marginal können Sätze angesehen werden, in denen das Definiendum schon am Satzbeginn von einem anderen Begriff abgegrenzt wird (vgl. Beispiel 92)):

92) Im Gegensatz zu den meist den Auftrieb eines Profils nutzenden Horizontalachsen-Windkraftanlagen ist bei den Windturbinen mit vertikaler Achse das Widerstandsprinzip vorherrschend.

Unter den durch eine Verbalphrase eingeleiteten KRCs verdienen die Sätze mit „es"-Konstruktion, worunter hier die beiden bereits diskutierten Verwendungsweisen (nicht-phorisches „es", vgl. Beispiel 93), und thetische Präsentative) fallen, als Sonderfälle ebenso Erwähnung wie der in den untersuchten Daten einzige Kontext mit satzeinleitendem „man" (vgl. Beispiel 58), Abschnitt 5.2.2.5.2).

93) Es wird bei Windkraftanlagen zwischen offshore und onshore Anlagen unterschieden.

Bei den anderen Beispielen in dieser Gruppe scheint es sich dagegen um diskursiv motivierte Abweichungen vom normalen Stellungsmodell zu handeln. Nicht zuletzt können KRCs auch durch Teilsätze eingeleitet werden – die Sätze 94) und 5 illustrieren unterschiedliche Typen von Teilsätzen.

94) Wird der Jahresertrag durch die Nennleistung geteilt, ergeben sich die sogenannten Jahresvolllaststunden.

95) Um die erheblich stärkeren Winde auf See nutzen zu können, wird in Deutschland vermehrt die Errichtung von Windparks auf dem offenen Meer, sogenannte Offshore-Windparks, geplant.

5.2.3.3.3 Zwischenfazit

Obige Ausführungen machen deutlich, dass in Bezug auf die grammatisch-syntaktische Realisierung des Definiendums, seine Thematisierung sowie die inhaltlich-strukturelle Ausgestaltung der KRCs eine beträchtliche Variationsbreite in den untersuchten Datensätzen sowie zwischen den Sprachen Russisch und Deutsch existiert. So wird zunächst deutlich, dass Definienda in den russischen KRCs häufiger als in den deutschen Daten als Subjekt realisiert und zudem auch häufiger direkt am Beginn des Satzes – als Subjekt oder präpositionales Objekt – positioniert werden, wobei jedoch beide Sprachen eine deutliche Tendenz zur Realisierung des Definiendums als satzinitiales Subjekt und damit als unmarkiertes Thema zeigen. Insbesondere letzterer Umstand könnte angesichts der im Russischen existierenden größeren syntaktischen Freiheit zunächst Erstaunen hervorrufen. Zu bedenken ist allerdings, dass das Russische aufgrund ebendieser Eigenschaft Positionsmerkmale als Mittel der inhaltlichen Akzentuierung von Satzbestandteilen (vgl. Bendixen/Voigt 1999: 213) nutzt, wobei gerade Satzbeginn und –ende markierte Positionen darstellen. Aus der beschriebenen Tendenz russischer KRCs zur satzinitialen Positionierung des Definiendums könnte mithin möglicherweise die Absicht einer Betonung ebendieses Definiendums herausgelesen werden, jedoch können endgültige Aussagen erst aufgrund umfassenderer Untersuchungen formuliert werden.
Im Vergleich zum Russischen mögen im Falle des Deutschen auch gewisse syntaktische Regularitäten (z. B. Inversion der Satzglieder bei Verwendung bestimmter Konjunktionen oder bei Satzeinleitung durch eine Adverbialbestimmung) sowie die Tendenz zur Thematisierung des Definiendums nicht als Subjekt, sondern als Situativkomplement zu den beobachteten intersprachlichen Unterschieden beigetragen haben. Eine hervorgehobene Realisierungsart des Definiendums ist im Deutschen zudem

die als – häufig ebenfalls satzinitiales – Situativkomplement: Wie bereits dargestellt, führt beispielsweise Cramer (2011) die deutsche Präposition „bei“, mit der solche Situativkomplemente häufig gebildet werden, als eigenes Extraktionsmuster an, womit auch in Bezug auf dieses Merkmal eine Korrespondenz zwischen Literatur und empirischen Daten hergestellt werden konnte.

5.2.3.4 Syntaktische Komplexität

Der simple Algorithmus für das Auslesen der syntaktischen Komplexität der deutschen KRCs zählt die durch den *parser* am Beginn jeder Zeile ausgegebenen Leerzeichen. Diese entsprechen selbstredend keinem etablierten Maß für die Beurteilung der syntaktischen Komplexität von Sätzen, allerdings kann aus diesen Informationen abgeleitet werden, wieviele übergeordnete Knoten im Syntaxbaum über der am tiefsten eingebetteten Konstituente existieren. Zum Beispiel gibt der Algorithmus für den in Abbildung 19 (Abschnit 5.2.3.1) dargestellten KRC den Wert *y=10* aus. Dieser trifft auf die Konstituenten *„(PP (APPR in) (ART der) (ADJA offenen) (NN See))“*, *„(PP (APPR vor) (ART der) (NN Küste))“* und *„(VVPP errichtet))“* gleichermaßen zu. Die Zahl *z* der übergeordneten Knoten errechnet sich nun mittels der Formel $z = \frac{y}{2}$, sie entspricht für das gegebene Beispiel also dem Wert *z = 5*. In der graphischen Ausgabe des *parsers* sind diese übergeordneten Knoten jeweils schwächer eingerückt als die am tiefsten eingebetteten Knoten. Die Vektoren der für jeden Satz ausgegebenen Komplexitätszahlen können statistisch miteinander verglichen werden, um globale Aussagen in Bezug auf die syntaktische Komplexität komplementärer Datensätze zu formulieren.
Als Signifikanztest zur Bewertung der Unterschiede zwischen den Korpora kann der für den Mittelwertvergleich von Zahlenreihen üblicherweise verwendete T-Test nicht angewandt werden, da er eine Normalverteilung des beobachteten Merkmals und Varianzgleichheit zwischen den untersuchten Datensätzen voraussetzt (Manning/Schütze 1999: 163). Für den größeren Teil der untersuchten Daten sind diese Annahmen nicht erfüllt, so dass für den Vergleich der Daten mit dem Mann-Whitney-U-Test ein Testverfahren verwendet wird, das als nicht-parametrisches Gegenstück des T-Tests gilt. Dieser Test prüft zwei unabhängige Stichproben darauf, ob sie gleiche Verteilungen aufweisen oder nicht (Sachs/Hedderich 2009[13]: 452). Die Null- und Alternativhypothese lauten entsprechend:

- H_0: $F_1(x) = F_2(x)$ – die Verteilungsfunktionen der Stichproben sind gleich
- H_1: $F_1(x) \neq F_2(x)$ – die Verteilungsfunktionen der Stichproben sind unterschiedlich

Tabelle 56 fasst nun die Ergebnisse der statistischen Auswertung der maximalen Einbettungstiefe für die 632 KRCs des deutschen Goldstandards sowie 632 zufällig ausgewählte generische Sätze aus den Extraktionskorpora Sub-EKD und Sub-TKD und weitere 632 Sätze aus dem NCL zusammen. Dafür werden neben dem arithmetischen Mittel für jeden Datensatz der Median, der denjenigen Wert bezeichnet, der die Anzahl der Beobachtungen in zwei gleich große Hälften teilt (ebd.: 66), das α-gestutzte Mittel, das als robuster Mittelwert gilt (ebd.: 75), die Standardabweichung, die Varianz (ebd.: 77) und die Ergebnisse (Prüfgröße und p-Wert) des Mann-Whitney-U-Tests angegeben.

	KRCs Goldstandard	Extraktionskorpora	NCL
Median	5	5	4
Arithmetisches Mittel	5,1171	4,7943	4,2263
α-gestutztes Mittel (α = 0,1)	4,9526	4,6581	4,1146
Standardabweichung	1,7603	1,7082	1,6513
Varianz	3,0988	2,9180	2,7268
Goldstandard-Extraktionskorpus:		**Goldstandard-NCL:**	
Prüfgröße	219.726,5		258.582,5
p	0,0017		< 0,0001

Tabelle 56: Analyse der maximalen syntaktischen Einbettungstiefe für deutsche KRCs im Vergleich zu Referenzdaten

Die Tabelle zeigt aufgrund der Ergebnisse des U-Tests sowohl für den Vergleich Goldstandard-Extraktionskorpus als auch den Vergleich Goldstandard-NCL statistisch signifikante Unterschiede zwischen den miteinander verglichenen Zahlenreihen auf. Von allen drei Datensätzen scheint dabei der Goldstandard die größte Variationsbreite aufzuweisen (größte Standardabweichung und größte Varianz). Allerdings bleiben die Unterschiede zwischen den jeweiligen Mittelwerten insbesondere für den Vergleich Goldstandard-Extraktionskorpus numerisch relativ klein. In Anbetracht der großen Standardabweichung des Goldstandards (rund 2 übergeordnete Konstitutenten bei einer durchschnittlichen Einbettungstiefe von rund 5) sowie der insgesamt unvollständigen konzeptuellen Modellierung der syntaktischen Komplexität erscheint das hier untersuchte Merkmal mithin als ein eher unsicheres, das auf relativ geringen Unterschieden basiert, dabei aber aufgrund des erforderlichen *parsing*-Prozesses mit hohen Annotationskosten einhergeht.

5.2.3.5 Satzlänge

5.2.3.5.1 Russisch

Tabelle 57 präsentiert die Ergebnisse der Satzlängenanalyse für die 560 KRCs des russischen Goldstandards im Vergleich zu den entsprechenden Referenzdatensätzen. Für diese ebenso wie für die nachfolgend beschriebene deutsche Analyse wurden nicht-alphanumerische Zeichen aus den Datensätzen entfernt, es wurde also pro Satz nur die Anzahl der Wörter und Zahlen gemessen.

	KRCs Goldstandard	Extraktionskorpora	RIC
Median	17	14	14
Arithmetisches Mittel	19,3768	16,6054	16,7464
α-gestutztes Mittel (α = 0,1)	17,9487	15,4375	14,9286
Standardabweichung	9,9713	9,3932	13,6136
Varianz	99,4267	88,2322	185,3309
Goldstandard-Extraktionskorpus:		**Goldstandard-RIC:**	
Prüfgröße	188.550		192.524,5
p	< 0,0001		< 0,0001

Tabelle 57: Analyse der Satzlänge für russische KRCs im Vergleich zu Referenzdaten

Aus der Tabelle geht hervor, dass die KRCs des russischen Goldstandards im Durchschnitt 19 Wörter umfassen, was deutlich länger ist als in beiden Vergleichsdatensätzen (in denen die Sätze im Schnitt rund 17 Wörter lang sind). Die Variation in allen drei Datensätzen ist beträchtlich, jedoch am höchsten im RIC (jeweils höchste Standardabweichung und Varianz). Der statistische Hypothesentest zeigt für beide Vergleiche jeweils hochsignifikante Verteilungsunterschiede.

5.2.3.5.2 Deutsch

Tabelle 58 präsentiert schließlich auch die Ergebnisse für die deutsche Satzlängenanalyse.

	KRCs Goldstandard	**Extraktionskorpora**	**NCL**
Median	18	16	14
Arithmetisches Mittel	19,3354	18,5949	14,9510
α-gestutztes Mittel (α = 0,1)	18,2885	17,4447	14,2589
Standardabweichung	8,9529	9,7970	7,2572
Varianz	80,1535	95,9815	52,668
Goldstandard-Extraktionskorpus:		**Goldstandard-NCL:**	
Prüfgröße	216.122,5		262.673
p	0,0114		< 0,0001

Tabelle 58: Analyse der Satzlänge für deutsche KRCs im Vergleich zu Referenzdaten

Auch hier zeigt der statistische Test wiederum Verteilungsunterschiede für beide Vergleiche. Das Ergebnis bestätigt zudem insofern die Ausführungen zur syntaktischen Komplexität, als der Goldstandard hier wie dort die höchsten Durchschnittswerte erreicht und auch die Länge des Satzes üblicherweise als Komplexitätsfaktor bewertet wird. Andererseits kann auch bei den hier untersuchten Datensätzen wieder ein erhebliches Variationsspektrum beobachtet werden (Standardabweichung von 9 Wörtern bei einer durchschnittlichen Satzlänge von rund 19 Wörtern für die deutschen KRCs). Anders als im Falle des Russischen weisen die Extraktionskorpora dabei die stärkste Variation auf. Im Vergleich zu diesem Variationsspektrum erscheinen die Mittelwert-Unterschiede zumindest zwischen dem Goldstandard und der Stichprobe aus den Extraktionskorpora wiederum als numerisch geringfügig.

5.2.3.5.3 Zwischenfazit

Die starke Variation des Merkmals Satzlänge in allen untersuchten Datensätzen bei zumindest für das Deutsche geringen numerischen Mittelwertunterschieden zwischen Goldstandard und Extraktionskorpus lässt das Merkmal trotz der signifikanten Verteilungsunterschiede wiederum als unsicher erscheinen. Dies bedeutet, dass trotz der statistisch signifikanten Unterschiede Zweifel daran bestehen, dass dieses Merkmal tatsächlich dabei hilft, KRCs in den Extraktionskorpora von Nicht-KRCs zu unterscheiden.

5.2.3.6 Phraseologie

5.2.3.6.1 Russisch

Tabelle 59 stellt die Ergebnisse für die Analyse der russischen Präpositionalphrasen aus den 560 KRCs des Goldstandards im Vergleich zu der Stichprobe generischer Sätze aus den Extraktionskorpora dar. Die Daten sind dabei nicht lemmatisiert, es wird also zwischen Groß- und Kleinschreibung unterschieden, die Unterscheidung zwischen terminologischen – also textspezifischen – und allgemein interessanten Einheiten wurde manuell vorgenommen. *n* bezeichnet die Anzahl der pro Korpus insgesamt extrahierten Einheiten.

Phrase	**Übersetzung**	**Frequenz KRCs Goldstandard (n = 1369)**	**Frequenz Extraktionskorpora (n = 1669)**
в зависимости*	in Abhängigkeit (von)	15	4
с помощью	mit Hilfe (von)	10	11
в соответствии	gemäß, in Übereinstimmung (mit)	6	7
при помощи*	unter Zuhilfename (von)	6	1
за счет	durch, aufgrund (von)	5	9
в результате	Im Ergebnis	4	6
с использованием*	unter Anwendung (von)	4	0
на основании*	auf Grundlage (von); am Boden (von)	4	0
с целью	mit dem Ziel	4	2
на основе	auf Grundlage (von)	4	4
в случае	im Falle (von)	4	1
В зависимости	In Abhängigkeit (von)	3	1
в большинстве	in der Mehrzahl	3	1
По способу	Nach der Art und Weise	3	0
в виде	in Gestalt (von)	3	3
для обеспечения	um sicherzustellen, dass	3	0
для определения	zur Bestimmung (von)	3	4
для изготовления	zur Herstellung (von)	3	0
в течение	im (Ver-)Lauf	3	4
Для создания	Zur Erzeugung (von)	2	0
по роду	nach der Art	2	0
по способу	nach der Art und Weise	2	0

в тех случаях	in diesen Fällen	2	0
в свою очередь	seinerseits	2	1
для выполнения	zur Ausführung	2	0
в качестве	(in der Eigenschaft) als	2	2
в том числе	darunter, einschließlich	2	3
По виду	Nach der Art	2	0
при эксплуатации	beim Betrieb	2	0
по величине	nach der Größe	2	0
По назначению	Nach ihrer Bestimmung	2	0
по форме	nach der Form	2	0
под действием	unter Einwirkung (von)	2	0
по типу	nach dem Typ	2	0
В качестве	(in der Eigenschaft) Als	2	0
для производства	für die Herstellung	2	1
в основном	im Wesentlichen	1	2
в частности	insbesondere	1	2
при работе	bei arbeitendem …	1	5
со стороны	seitens, von Seiten, auf der Seite …	0	2
Таким образом	Auf diese Weise	0	3
в настоящее время	gegenwärtig, heutzutage	0	3
в среднем	im Durchschnitt	0	2
с учетом	unter Berücksichtigung (von)	0	2
Кроме того	Außerdem	0	2
том числе	darunter, einschließlich	0	3
при необходимости	falls erforderlich	0	3
для того	um zu	0	3
в том	darin	0	4
При этом	Dabei	0	4

Tabelle 59: Häufigkeitsvergleich automatisch extrahierter Präpositionalphrasen für den russischen Goldstandard und die Zufallsstichprobe aus den Extraktionskorpora

Die Tabelle enthält zwar keine der gesuchten Geltungsrestriktionen, wohl aber eine Reihe funktional unterschiedlicher Wortverbindungen, z. B. generalisierende Wendungen wie *в большинстве* („in der Mehrzahl“), *в среднем* („im Durchschnitt, im Schnitt“), *в основном* („im Wesentlichen“), Zeitangaben wie *в настоящее время* („gegenwärtig, heutzutage“), konkretisierende Wendungen wie *в частности* („insbesondere“), Wendungen, die Klassifikationsmerkmale einführen, wie die Phrasen auf *по* und Wendungen mit kohäsionsbildender Funktion wie *таким образом* („auf diese Weise, so“). Mengenmäßig sind die untersuchten Datensätze ganz offenbar zu klein für eine quantitative Analyse, so zeigen nur die Phrasen *в зависимости* („in Abhängigkeit von“), *при*

помощи („unter Zuhilfenahme von“), *с использованием* („unter Verwendung“), *на основании* („auf Grundlage von“) und – wenn satzinitiale und innerhalb des Satzes angeordnete Varianten gemeinsam gezählt werden – *по способу* („nach der Art und Weise“) statistisch signifikante (nach dem Fisher-Test) Verteilungsunterschiede (statistische Signifikanz wird durch * markiert), die sich allerdings auch auf einem insgesamt sehr niedrigen Niveau bewegen und darüber hinaus auch von der Güte des verwendeten Extraktionsalgorithmus‘ abhängen, die hier nicht evaluiert wurde. Obwohl sowohl Pearson (1998) als auch Walter (2010) und Cramer (2011) auf Geltungsrestriktionen bzw. „geltungsregulierende Elemente“ wie „in der Regel“ (Walter 2010: 70) verweisen, zeigt sich, dass der Goldstandard zwar auf phraseologischer Ebene Einheiten enthält, welche die funktionalen Selektionskriterien widerzuspiegeln scheinen (z. B. Phrasen zur Einführung klassifikatorisch relevanter Merkmale oder zur Angabe von Funktionen), jedoch selbst einige Wendungen mit verallgemeinernder Bedeutung zumindest in den hier untersuchten kleinen Datensätzen eher für die Referenzdaten typisch zu sein scheinen. Die Verteilung beispielsweise von kohäsionsbildenden, insbesondere anaphorischen, Wendungen und Zeitangaben ist dagegen mit der bisherigen Darstellung der linguistischen Eigenschaften wissenshaltiger Kontexte kohärent.

5.2.3.6.2 Deutsch

Tabelle 60 präsentiert schließlich auch die Ergebnisse für die KRCs des deutschen Goldstandards im Vergleich zu generischen Sätzen aus den Extraktionskorpora.

Phrase	**Frequenz KRCs Goldstandard (n = 1558)**	**Frequenz Extraktionskorpora (n = 1432)**
zur Stromerzeugung*	14	0
zum Beispiel	12	4
zum Antrieb	7	2
zur Verfügung	6	2
im Gegensatz	5	1
im Regelfall	5	0
zum Heizen	5	0
im Wesentlichen	5	1
vor allem	5	3
mit Hilfe	4	1
in Form	4	0
zum Einsatz	3	4
zur Erzeugung	3	0
unter anderem	3	2
in der Regel	2	5
unter Umständen	2	2
in der Lage	2	2
zur direkten Nutzung	2	1

Tabelle 60: Häufigkeitsvergleich automatisch extrahierter Präpositionalphrasen für den deutschen Goldstandard und die Zufallsstichprobe aus den Extraktionskorpora

Für das Deutsche konnten mit Hilfe der beschriebenen Methode deutlich weniger relevante Präpositionalphrasen identifiziert werden. Auch sie scheinen zudem insofern mehrheitlich die funktionalen Selektionskriterien für KRCs widerzuspiegeln, als sie Funktionen benennen (z. B. „zum Antrieb“) oder bei der Darstellung von Fähigkeiten oder Eigenschaften eines bestimmten

Mechanismus‘ Anwendung finden. Darüber hinaus finden sich auch Phrasen der inhaltlichen Abgrenzung („im Gegensatz“) und verallgemeinernde bzw. geltungsregulierende Wendungen wie „vor allem“ und „in der Regel“, allerdings kann in Bezug auf deren Verteilung aufgrund der Datenlage keinerlei Verteilungstendenz angegeben werden, zumal auch Wendungen, die Unsicherheit ausdrücken („unter Umständen“), in den untersuchten Daten auftauchen. Statistisch signifikante Verteilungsunterschiede können für keine einzige der Phrasen in der Tabelle gezeigt werden. Die in der Literatur beschriebenen Geltungsrestriktionen bzw. Fachgebietsangaben können nicht gefunden werden.

5.2.3.6.3 Zwischenfazit

Im Ergebnis können die Darstellungen aus der zitierten Literatur nur ansatzweise bestätigt werden: Zwar wurden sowohl für das Russische als auch für das Deutsche „geltungsregulierende“ Phrasen im Sinne Walters (2010) gefunden, ihre Typizität für KRCs konnte jedoch keineswegs zweifelsfrei gezeigt werden. Neben geltungsregulierenden Elementen umfassen die extrahierten Präpositionalphrasen auch solche Wendungen, welche die Selektionskriterien widerspiegeln, also beispielsweise zum Ausdruck der ausgewählten semantischen Relationen dienen, sowie textlinguistische Steuerphrasen, wobei auch für diese Phrasen statistisch signifikante Verteilungsunterschiede nur in wenigen Fällen gezeigt werden konnten. Es ergibt sich folglich wiederum die Notwendigkeit von Studien auf Grundlage größerer Datensammlungen. Nicht zuletzt kann auch angesichts der geringen Frequenzen der identifizierten Wendungen deren Relevanz für die Extraktion wissenshaltiger Kontexte angezweifelt werden – es scheint sich erneut um ein Merkmal mit nur schwachen Ausprägungen individueller Werte zu handeln.

5.2.4 Zusammenfassung

In den obigen Abschnitten wurde mit qualitativen und quantitativen Mitteln eine sehr umfangreiche linguistische Analyse wissenshaltiger Kontexte durchgeführt, wie sie in dieser Detailliertheit in der referierten Literatur nicht anzutreffen ist. Aufgrund der so erhobenen Daten können nunmehr zwei Fragen diskutiert werden, die sich direkt aus der in Kapitel 3 referierten Literatur und der daraus erwachsenden Frage nach den linguistischen Eigenschaften wissenshaltiger Kontexte ergeben, und zwar:

- Sind wissenshaltige Kontexte durch einen eigenen Funktionalstil des Erklärens und Definierens gekennzeichnet bzw. können sie als eigene *sublanguage* oder Textsorte charakterisiert werden?
- Welche Gemeinsamkeiten und Unterschiede weisen wissenshaltige Kontexte im Deutschen und Russischen auf?

In Bezug auf die Frage nach einem möglichen Funktionalstil sind die ermittelten Ergebnisse uneindeutig. Zwar zeigen die untersuchten KRCs in beiden Sprachen gewisse Regularitäten auf den Ebenen des Lexikons, der Grammatik und auch der Syntax, allerdings kann auf allen Ebenen auch eine erhebliche Variation beobachtet werden. Problematisch ist zudem der Umstand, dass zwar Merkmale beobachtet werden konnten, die in der zitierten Literatur als Merkmale verallgemeinernder Aussagen angeführt werden, nun aber gerade aufgrund dieser Allgemeinheit wissenshaltiger Kontexte die Zuschreibung eines eigenen Funktionalstils kaum möglich scheint (hiermit sind insbesondere linguistisch unmarkierte Merkmale gemeint, z. B. das Tempus Präsens

oder die satzinitiale Position des Subjekts). Zu bedenken ist außerdem, dass – hierauf wurde im Lauf der Untersuchung mehrfach hingewiesen – die untersuchten Datensätze zumindest aus statistischer Sicht immer noch zu klein für stabile Schlussfolgerungen sind. Darüber hinaus wurden ganz wesentliche, nämlich textlinguistische Aspekte von Funktionalstilen in Bezug auf wissenshaltige Kontexte in dieser Untersuchung überhaupt nicht thematisiert. Nicht zuletzt kann trotz mehrfacher Annotation aufgrund der Art und Weise der Erstellung des Goldstandards eine Verzerrung der Merkmale der wissenshaltigen Kontexte in den untersuchten Stichproben im Verhältnis zu der nicht untersuchbaren Grundgesamtheit aller deutschen und russischen KRCs nicht vollständig ausgeschlossen werden. Andererseits wurde aber auch deutlich, dass zumindest die definitorischen Aussagen in der Literatur häufig zugeschriebenen lexikalischen und phraseologischen Merkmale eher selten auftreten und nur in wenigen Fällen tatsächlich insofern diskriminativ wirken können, als sie spezifische Verteilungen im Vergleich zu Referenzdatensätzen aufweisen: Statt von Merkmalen sollte hier eher von Tendenzen gesprochen werden, deren Nützlichkeit für praktische Extraktionsaufgaben noch zu zeigen ist. Im statistischen Sinne relevantere Unterschiede zwischen KRCs und generischen Sätzen konnten dagegen auf grammatischer Ebene gezeigt werden, wo bestimmte sprachliche Strukturen bzw. das Fehlen bestimmter Elemente auf das Vorliegen eines KRCs hindeuten können. Statt einer pauschalen Subsumption von KRCs unter einen eigenen Funktionalstil erscheint daher die Strategie sinnvoller, KRCs anhand einer finiten Menge von Mikrofunktionen – die in Kapitel 4 referierten Validitätskriterien könnten hierfür als Anhaltspunkte dienen – zu charakterisieren, deren Vorliegen durch bestimmte sprachliche Signale kenntlich gemacht wird, die allerdings unter bestimmten Umständen auch generischen Sätzen zugeordnet werden können.

Die Schwierigkeit einer trennscharfen Abgrenzung wissenshaltiger Kontexte von generischen Sätzen scheint nicht zuletzt auch die auf der Ebene der Humanannotation (Kapitel 4 sowie die theoriebasierte Diskussion in Kapitel 3) beobachteten Schwierigkeiten widerzuspiegeln. Ebenso wie aus der Literatur hervorgeht, dass Annotatoren bestimmte Typen definitorischer Aussagen – etwa aristotelische Definitionen – bevorzugen, wofür auch die durchgeführte Annotationsstudie zumindest Indizien liefert, lassen sich auf linguistischer Ebene Tendenzen – An- und Abwesenheiten bestimmter linguistischer Formen und Strukturen – beobachten, eine vollständige Abgrenzung der Klassen KRC vs. generischer Satz scheint zumindest auf Grundlage der bisher referierten Ergebnisse jedoch nicht möglich.

Die Validität der ermittelten Ergebnisse – und damit auch die beschriebenen Tendenzen in der linguistischen Gestaltung wissenshaltiger Kontexte – stützt allerdings die Tatsache, dass die russischen und deutschen KRCs eine Reihe relevanter Gemeinsamkeiten aufweisen. Auf die Ähnlichkeit der in der Literatur für eine Reihe unterschiedlicher Sprachen beschriebenen Extraktionsmuster war bereits hingewiesen worden. Die im Rahmen dieser Analyse ermittelten Ähnlichkeiten von KRCs erstrecken sich – wenn von systembedingten, sprachenpaarspezifischen Unterschieden wie etwa dem Fehlen des Verbalaspekts im Deutschen abgesehen wird – dabei auf die Verteilung bestimmter Wortarten, potentielle lexikalische Bausteine für Extraktionsmuster, die Realisierung und Thematisierung des Definiendums sowie semantische und grammatische Merkmale der Verben.

5.3 Multivariate Analyse deutscher KRCs

5.3.1 Einführung

Bislang wurden uni- und bivariate statistische sowie qualitative (manuelle) Analyseverfahren eingesetzt, um theoriebasierte Hypothesen zu linguistischen Merkmalen wissenshaltiger Kontexte zu erhärten oder zu widerlegen. Mit Hilfe dieser Analysen wurden einzelne Merkmale isoliert von der Gesamtheit aller möglichen Merkmale wissenshaltiger Kontexte daraufhin untersucht, ob sie in Bezug auf die interessierenden Datensätze (KRCs und Vergleichsdaten) signifikante, d. h. unter bestimmten Bedingungen reproduzierbare Verteilungsunterschiede aufweisen und ob ihnen mithin in Bezug auf die Klassifikation eines Satzes als KRC oder als generischer Satz ein diskriminatorisches Potential unterstellt werden kann.
Die Analysen ergaben lexikalische sowie strukturell-grammatische Unterschiede zwischen KRCs und generischen Sätzen. So konnten auf lexikalischer Ebene Unterschiede in der Verteilung bestimmter Wortarten und Lemmata einschließlich *knowledge patterns* ebenso ermittelt werden wie Differenzen in der Wahl semantischer Klassen für Prädikate und in der lexikalischen Variabilität dieser Prädikate. Auf strukturell-grammatischer Ebene wurden dagegen Besonderheiten wissenshaltiger Kontexte in Bezug auf Verbformen (z. B. Tempus und Modus), Strukturen (z. B. unbestimmt-persönliche Sätze im Russischen), die Thematisierung des Definiendums (z. B. Verwendung von Situativkomplementen im Deutschen) aufgezeigt. Diese Ergebnisse können zur Differenzierung der in der Literatur genannten Hypothesen und nicht zuletzt auch für Extraktionsexperimente genutzt werden. Allerdings weisen solche Einzelanalysen auch eine Reihe von Nachteilen auf, und zwar:

- Sie erlauben keine ganzheitliche Betrachtung der Daten und geben insbesondere keine Einblicke in deren innere Struktur.
- Insbesondere können also die linguistischen Merkmale nicht nach ihrer Wichtigkeit oder diskriminatorischen Kraft in Bezug auf das Klassifikationsproblem KRC vs. generischer Satz geordnet werden.
- Obwohl in einigen Fällen Konfidenzintervalle genutzt wurden, um statistisch signifikante Verteilungsunterschiede auf ihre Stärke, d. h. letztlich auf ihre Relevanz für das formulierte Klassifikationsproblem zu untersuchen, bleiben diese Ergebnisse dennoch eher abstrakt.

Aus diesen Gründen soll nun mit der logistischen Regression ein multivariates Analyseverfahren eingesetzt werden, um zu untersuchen, welche Leistung die ermittelten Merkmale wissenshaltiger Kontexte in Bezug auf die genannte Klassifikationsentscheidung im Ensemble erbringen. Konkret sollen diejenigen Merkmale, die sich im Rahmen der bisherigen Analysen als charakteristisch für KRCs erwiesen haben, daraufhin überprüft werden, ob sie – einzeln oder im Zusammenspiel mit anderen Merkmalen – mit einer akzeptablen Genauigkeit wissenshaltige Kontexte von generischen Sätzen unterscheiden. Des Weiteren kann ein solches Modell zur Klassifizierung ungesehener Daten und damit als einfache Methode für die automatische Identifikation wissenshaltiger Kontexte genutzt werden. Die Analyse wird dabei nur für die deutschen Daten durchgeführt, da sie an die automatische Extraktion der interessierenden Merkmale aus ausreichend großen linguistischen Ressourcen (z. B. GermaNet) gebunden ist, die für das Russische nicht in jedem Fall zur Verfügung stehen.

5.3.2 Logistische Regression

Die logistische Regression ist ein klassisches Verfahren der statistischen Modellbildung, mit dem die Verteilung einer dichotomomen Zielgröße (z. B. KRC vs. generischer Satz) beschrieben werden kann (Sachs/Hedderich 2009[13]: 675 ff.). Modelliert wird in diesem Fall die unbekannte bedingte Erfolgswahrscheinlichkeit $\pi(x) = P(Y = 1|X = x)$ auf Grundlage der zur Modellbildung verwendeten unabhängigen Variablen bzw. *features,* d. h. die Wahrscheinlichkeit, dass die Zielgröße $y \epsilon Y$ – gegeben eine bestimmte Wertebelegung des Vektors der unabhängigen Variablen $x \epsilon X$ – einen bestimmten Wert (aus zwei möglichen) annimmt. Da eine Wahrscheinlichkeit berechnet wird, unterscheidet sich die Modellbildung bei der logistischen Regression im Vergleich zu einem linearen Regressionsmodell dabei dadurch, dass die Linearkombination der einzelnen Merkmale (*features*) durch eine geeignete Transformation derart angepasst werden muss, dass gilt: $0 \leq \pi \leq 1$. Dies wird durch die sogenannte logistische Transformation erreicht:

96) $$\pi(x) = \frac{e^{\sum_{i=1}^{n} w_i x_i + b}}{1 + e^{\sum_{i=1}^{n} w_i x_i + b}}$$

In dieser Darstellung bezeichnet x_i das i-te zur Modellbildung verwandte Merkmal, während w_i sich auf dessen Koeffizienten bezieht. Die Koeffizienten beschreiben die Stärke des Einflusses eines Merkmals auf die Zielgröße, also beispielsweise darauf, ob ein Satz als KRC oder aber als generische Aussage klassifiziert wird. Der Wert *b* beschreibt dagegen eine Zufallskomponente (ebd.: 649).

5.3.3 Merkmale und Daten

Die linguistischen Merkmale, auf deren Grundlage das Regressionsmodell erstellt werden soll, können für jeden zu untersuchen deutschen Satz automatisch aus den in Kapitel 4 beschriebenen Goldstandard-Korpora im XML-Format extrahiert werden. Dazu werden XSLT-Regeln ebenso verwendet wie die Perl- und Java-APIs für GermaNet und einfache Perl-Skripte. Die Merkmalsauswahl stützt sich dabei auf die bisherigen linguistischen Analysen und insbesondere auf diejenigen Merkmale, für die spezifische Verteilungen im Ergebnis aller durchgeführten Vergleiche gefunden wurden. Tabelle 61 gibt einen Überblick über die extrahierten Merkmale.

Merkmalstyp	**Ausprägung**	**Datentyp**
Allgemein	Satz-ID	numerisch
	Satzlänge	numerisch
Zielvariable	KRC	binär
Boost- und Stopwörter, typische Wortarten	Anzahl der Nomina (*tag* NN)	numerisch
	Anteil der Nomina	numerisch
	Anzahl attributiver Adjektive (*tag* ADJA)	numerisch
	Anteil attributiver Adjektive	numerisch
	Anzahl koordinierender Konjunktionen (*tag* KON)	numerisch
	Anteil koordinierender Konjunktionen	numerisch
	Anzahl der Adverbien (*tag* ADV)	numerisch
	Anteil der Adverbien	numerisch
	Anzahl prädikativer Adjektive (*tag* ADJD)	numerisch
	Anteil prädikativer Adjektive	numerisch
	Anzahl von Verben im Partizip II (*tag* VVPP)	numerisch

	Anteil von Verben im Partizip II	numerisch
	Anzahl von Eigennamen (*tag* NE)	numerisch
	Anteil an Eigennamen	numerisch
	Anzahl attributiver Indefinitpronomen (*tag* PIAT)	numerisch
	Anteil attributiver Indefinitpronomen	numerisch
	Anzahl finiter Hilfsverben (*tag* VMFIN)	numerisch
	Anteil finiter Hilfsverben	numerisch
	Anzahl von Infinitiven (*tag* VVINF)	numerisch
	Anteil von Infinitiven	numerisch
	Anzahl von Personalpronomen (*tag* PPER)	numerisch
	Anteil von Personalpronomen	numerisch
	Anzahl von Negationspartikeln (*tag* PTKNEG)	numerisch
	Anteil von Negationspartikeln	numerisch
	Anzahl nebenordnender Konjunktionen (*tag* KOUS)	numerisch
	Anteil nebenordnender Konjunktionen	numerisch
	Anzahl der Possessivpronomen (*tag* PPOSAT)	numerisch
	Anteil der Possessivpronomen	numerisch
	Anzahl von Boostwörtern (Lemmata „entweder“, „oder“, „Begriff“, „Form“, „welche“, „all“, „mehrer“, „jed“, „grundsätzlich“, „meist“, „grundprinzip“, „grundgedanke“)	numerisch
	Anteil von Boostwörtern	numerisch
	Anzahl von Stopwörtern (Lemmata „erst“, „nur“, „jetzt“, „gut“, „aber“, „dass“, „nicht“, „für“, „nach“, „dies“, „er“, „es“ (in nicht-satzinitialer Position), „ich“, „ihr“, „Sie“, „sie“, „müssen“, „keine“, „wenige“, „je“, „stark“, „viele“, „folgend“, „zeigen“, „allerdings“, „gleich“, „weiter“, „beide“, „erreichen“, „bis“, „arbeiten“, „pro“, „liegen“)	numerisch
	Anteil von Stopwörtern	numerisch
	Anteil Boostwörter minus Anteil Stopwörter	numerisch
Verbsemantik	Anzahl der GermaNet-Klassen des Prädikats	numerisch
	Anzahl der GermaNet-Synsets, in denen das Prädikat erscheint	numerisch
Grammatische und strukturelle Merkmale	Anzahl der als Prädikat annotierten Einheiten	numerisch
	Anteil der als Prädikat annotierten Einheiten	numerisch
	Anzahl von Verben mit KRC-typischer Morphologie (Indikativ, 3. Person, Präsens, Plural)	numerisch
	Anteil von Verben mit KRC-typischer Morphologie	numerisch
	Anzahl von Verben mit KRC-untypischer Morphologie (Konjunktiv, 1. Person, Vergangenheit)	numerisch
	Anteil von Verben mit KRC-untypischer Morphologie	numerisch
	Anteil von Verben mit KRC-typischer Morphologie minus Anteil von Verben mit KRC-untypischer Morphologie	numerisch
	Phorischer Satz (Satzbeginn: Es + konjugiertes Verb)	binär
Terminologische Merkmale	Anzahl terminologischer Nomina	numerisch
	Term in Subjektposition	binär
	Term als präpositionales Objekt	binär
	Term als präpositionales Attribut	binär
	Term als Genitivattribut	binär
	Term als direktes Objekt	binär
	Term in satzinitialer Position (erstes, zweites oder drittes *token*)	binär

Phraseologie	KRC-typische Phrase („zum Beispiel“, „im Gegensatz“, „im Regelfall“, „im Wesentlichen“, „vor allem“)	binär

Tabelle 61: Merkmale für die multivariate Analyse deutscher KRCs

Auf diese Weise kann jeder Satz als Merkmalsvektor kodiert werden. Tabelle 62 gibt ein Beispiel für den KRC: „Der Darrieus-Rotor ist eine Windenergieanlagenbauart mit vertikaler Rotationsachse.“ aus dem Sub-EKD.

Merkmal	**Erklärung**	**Wert**
ID	Satz-ID	109
KRC	KRC	1
SENT_L	Satzlänge	9
NOUNS_N	Anzahl der Nomina (*tag* NN)	3
NOUNS_PROP	Anteil der Nomina	0,3
ADJA_N	Anzahl attributiver Adjektive (*tag* ADJA)	1
ADJA_PROP	Anteil atributiver Adjektive	0,1
KON_N	Anzahl koordinierender Konjunktionen (*tag* KON)	0
KON_PROP	Anteil koordinierender Konjunktionen	0
ADV_N	Anzahl der Adverbien (*tag* ADV)	0
ADV_PROP	Anteil der Adverbien	0
ADJD_N	Anzahl prädikativer Adjektive (*tag* ADJD)	0
ADJD_PROP	Anteil prädikativer Adjektive	0
VVPP_N	Anzahl von Verben im Partizip II (*tag* VVPP)	0
VVPP_PROP	Anteil von Verben im Partizip II	0
NE_N	Anzahl von Eigennamen (*tag* NE)	0
NE_PROP	Anteil an Eigennamen	0
PIAT_N	Anzahl attributiver Indefinitpronomen (*tag* PIAT)	0
PIAT_PROP	Anteil attributiver Indefinitpronomen	0
VMFIN_N	Anzahl finiter Hilfsverben (*tag* VMFIN)	0
VMFIN_PROP	Anteil finiter Hilfsverben	0
VVINF_N	Anzahl von Infinitiven (*tag* VVINF)	0
VVINF_PROP	Anteil von Infinitiven	0
PPER_N	Anzahl von Personalpronomen (*tag* PPER)	0
PPER_PROP	Anteil von Personalpronomen	0
PTKNEG_N	Anzahl von Negationspartikeln (*tag* PTKNEG)	0
PTKNEG_PROP	Anteil von Negationspartikeln	0
KOUS_N	Anzahl nebenordnender Konjunktionen (*tag* KOUS)	0
KOUS_PROP	Anteil nebenordnener Konjunktionen	0
PPOSAT_N	Anzahl der Possessivpronomen (*tag* PPOSAT)	0
PPOSAT_PROP	Anteil der Possessivpronomen	0
BOOST_N	Anzahl von Boostwörtern	0
BOOST_PROP	Anteil von Boostwörtern	0
STOP_N	Anzahl von Stopwörtern	0
STOP_PROP	Anteil von Stopwörtern	0
BOOST_STOP_R	Differenz BOOST_PROP-STOP_PROP	0
PRED_CLASS_N	Anzahl der GermaNet-Klassen des Prädikats	4
PRED_N	Anzahl der als Prädikat annotierten Einheiten	1
PRED_R	Anteil der als Prädikat annotierten Einheiten	0,1
SYNSETS	Anzahl der GermaNet-Synsets, in denen das Prädikat erscheint	13

BOOSTM_N	Anzahl von Verben mit KRC-typischer Morphologie	1
BOOSTM_PROP	Anteil von Verben mit KRC-typischer Morphologie	0,1
STOPM_N	Anzahl von Verben mit KRC-untypischer Morphologie	0
STOPM_PROP	Anteil von Verben mit KRC-untypischer Morphologie	0
BOOSTM_STOPM_R	Differenz BOOSTM_PROP-STOPM_PROP	0,1
PHOR	Phorischer Satz	0
TERM_N	Anzahl terminologischer Nomina	1
TERM_S	Term in Subjektposition	1
TERM_PO	Term als präpositionales Objekt	0
TERM_PA	Term als präpositionales Attribut	0
TERM_GA	Term als Genitivattribut	0
TERM_DO	Term als direktes Objekt	0
TERM_INI	Term in satzinitialer Position	1
PHRASE	KRC-typische Phrase	0

Tabelle 62: Merkmalsvektor für einen Beispielsatz

In den XML-formatierten deutschen Sub-Korpora des Goldstandards wurden 626 (339+287; die Zahlen weichen von den Angaben des Goldstandards ab, da einerseits einige KRCs aufgrund falscher Zeilenumbrüche nicht als solche extrahiert werden können und andererseits einige wenige KRCs mit lediglich formalen Variationen mehrfach im Korpus auftreten, so dass sie durch die in Kapitel 4 durchgeführte Entfernung exakter Duplikate nicht erkannt wurden) extrahierbare KRCs anhand ihrer IDs identifiziert. Für die multivariate Modellierung der linguistischen Merkmale von KRCs im Vergleich zu generischen Sätzen unabhängig vom behandelten Fachgebiet wurden Sub-EKD und Sub-TKD miteinander vereint. Dieser Datensatz wurde nun mittels zufälliger Satzauswahl in einen balancierten, d. h. aus 563 KRCs und 563 generischen Sätzen bestehenden Trainings-Datensatz, der die Grundlage der Modellierung bildet, sowie einen Testdatensatz für die Experimente in den Abschnitten 5.3.4.3 und 5.3.4.4 aufgeteilt. Letzterer wurde ebenfalls durch die zufällige Auswahl von Sätzen aus den Daten gebildet und enthält 2100 generische Sätze sowie 62 KRCs, womit eine gute Annäherung an die Verteilung der KRCs in den Sub-Korpora des Goldstandards erreicht ist.

5.3.4 Modellierung und Ergebnisse

Die Modellierung wurde in R (R Development Core Team 2012) gemäß der Beschreibung in Sachs und Hedderich (2009[13]: 674 ff.) durchgeführt. Allerdings wurden absolute Frequenzen von Wortarten und Prädikaten nicht in die Modellbildung einbezogen, da in Bezug auf diese Korrelationen mit dem Merkmal „Satzlänge“ wahrscheinlich sind, die das errechnete Modell verzerren könnten. Des Weiteren wurden einige weitere Merkmale aus der Modellierung ausgeschlossen, da eine Korrelationsanalyse Zusammenhänge zwischen diesen ergeben hat (vgl. Wollschläger 2012[2]: 197)[126]. Das im Ergebnis errechnete Modell wurde einem Prozess der Merkmalsauswahl (s. nächsten Abschnitt) unterzogen.

R gibt für jedes Merkmal in einem Regressionsmodell neben dem errechneten Regressionskoeffizienten u. a. den Standardfehler und die sogenannte Wald-Statistik (Sachs/Hedderich 2009[13]: 678 f.) für diesen Koeffizienten sowie das Ergebnis eines Signifikanztests aus, d. h. eine Angabe, ob ein gegebenes Merkmal signifikant zur Anpassung des Modells beiträgt. Als Gütemaß für die Anpassung des Modells an die tatsächlich beobachteten Daten wird zusätzlich die

[126] Dies betraf insbesondere das Merkmal TERM_S, das mit mehreren anderen Variablen korrelierte.

Devianz ausgegeben, welche die Summe der Abweichungen des mathematisch angepassten Modells von den empirischen Daten bezeichnet. Ein auf Grundlage aller zur Verfügung stehenden Variablen errechnetes Modell erreicht eine Devianz von 712,01 im Vergleich zu einer Nulldevianz, in der nur die Zufallskomponente berücksichtigt wird (ebd.: 679 f.), von 1560,97. Ein Chi-Quadrat-Test auf Grundlage der Devianzdifferenzen für die einzelnen Parameter des Modells gemäß Sachs/Hedderich (ebd.) ergibt, dass die folgenden Variablen signifikant (auf einem Niveau von 0,05) zur Minderung der Devianz im Vergleich zum Nullmodell beitragen:

- SENT_L,
- NOUNS_PROP,
- ADJA_PROP,
- KON_PROP,
- ADV_PROP,
- VVPP_PROP,
- NE_PROP,
- PPER_PROP,
- KOUS_PROP,
- PPOSAT_PROP,
- BOOST_STOP_R,
- PRED_CLASS_N,
- BOOSTM_STOPM_R,
- TERM_N,
- TERM_PA
- TERM_GA,
- TERM_INI.

5.3.4.1 Variablenauswahl

Des Weiteren bietet R die Möglichkeit, eine Variablenauswahl gemäß dem AIC-Kriterium (AIC-Modell, ebd.: 685 f.) durchzuführen, mit dessen Hilfe automatisch ein Modell mit weniger abhängigen Variablen und geringer Devianz ermittelt werden kann. Dieses Modell erreicht eine Devianz von 717,64 und selektiert die folgenden Variablen (Signifikanz gemäß des Tests auf Grundlage der Devianzdifferenzen wird durch * gekennzeichnet):

- SENT_L*,
- ADJA_PROP*,
- ADV_PROP*,
- VVPP_PROP*,
- PIAT_PROP
- PPER_PROP*,
- KOUS_PROP*,
- BOOST_STOP_R*,
- PRED_CLASS_N*,
- BOOSTM_STOPM_R*,
- TERM_N*,
- TERM_PA*
- TERM_PO,
- TERM_GA*,
- TERM_DO,
- TERM_INI*.

5.3.4.2 Evaluation der Modellanpassung

Bevor das Modell interpretiert wird, ist dessen Anpassungsgüte zu evaluieren. Hierfür bietet R eine Reihe von Möglichkeiten. Hier sind insbesondere die Residuenanalyse und die Bewertung mit Pseudo-Bestimmtheitsmaßen zu nennen.
Die Residuen bezeichnen Abweichungen des berechneten Modells von der empirisch beobachtbaren Verteilung der Daten (ebd.: 651, 686) und sollten unter korrekten Modellannahmen annähernd normalverteilt sein. Abbildung 25 zeigt einen sogenannten QQ-Plot (ebd.: 387) der studentisierten Devianzresiduen (links) sowie ein Histogramm derselben (rechts) für das AIC-Modell. Beide Darstellungen vergleichen die Residuen mit einer Normalverteilung.

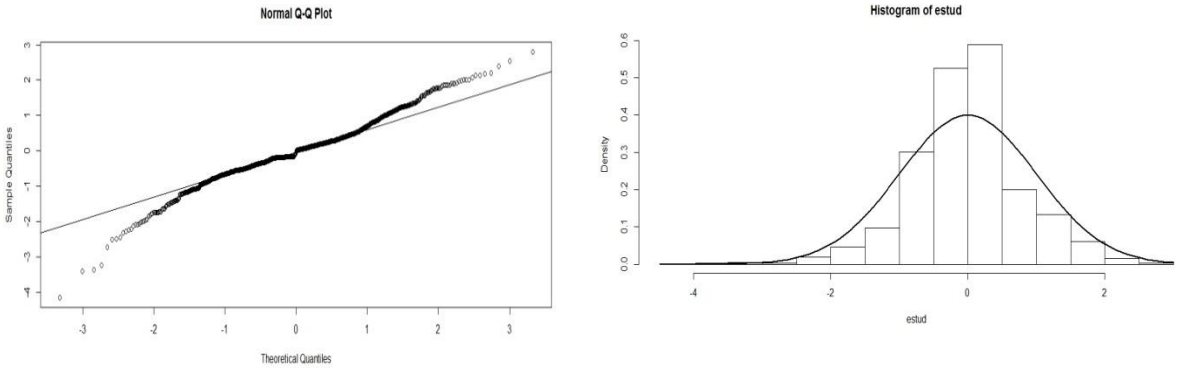


Abbildung 25: QQ-Plot und Histogramm der studentisierten Devianzresiduen für das AIC-Modell

Die Darstellung zeigt für beide Plots Abweichungen von der Normalverteilung mit allerdings eher schwach besetzten Verteilungsenden bzw. eher geringen Abweichungen, so dass von einer ausreichend guten Modellanpassung ausgegangen werden kann. Als Pseudo-Bestimmtheitsmaß kann in R u. a. Pseudo-R^2 nach McFadden berechnet werden (ebd.: 687). Für das finale Modell erreicht dieser Wert $R^2 = 0{,}5403$ und bezeichnet damit eine deutliche Verbesserung gegenüber dem Nullmodell. Eine 10-fache Kreuzevaluation auf den Trainingsdaten, in deren Verlauf das Modell für die Klassifikation auf den Trainingsdaten verwendet wird, um die Güte der Modellanpassung zu überprüfen, ergab einen Klassifikationsfehler von etwa 10%, d. h., wenn die Trainingsdaten in zehn Sub-Datensätze aufgeteilt werden, von denen in zehn Testdurchläufen jeweils neun für die Modellanpassung und ein Datensatz für die Vorhersage von Wahrscheinlichkeiten genutzt werden, ergeben sich im Schnitt 10% falsche Vorhersagen. Insgesamt wird daher von einer ausreichend guten Modellanpassung ausgegangen.

5.3.4.3 Interpretation der Regressionskoeffizienten

Für die hier untersuchte Fragestellung sind die Modellparameter zu interpretieren. Sie werden in Tabelle 63 gemeinsam mit den errechneten Regressionskoeffizienten sowie den dazugehörigen Standardfehlern und Konfidenzintervallen zusammengefasst. Negative Regressionskoeffizienten markieren dabei Merkmale, die für KRCs als untypisch anzusehen sind, wohingegen positive Vorzeichen auf einen positiven Zusammenhang zwischen dem betreffenden Merkmal und wissenshaltigen Kontexten verweisen.

Merkmal	Koeffizient	Standardfehler	KI
SENT_L	0,0254	0,0083	[0,0092, 0,0418]
ADJA_PROP	5,3203	1,2668	[2,8177, 7,7913]
ADV_PROP	-4,3401	1,7785	[-7,8874, -0,9097]
VVPP_PROP	13,4947	2,3619	[8,9015, 18,1735]
PIAT_PROP*	-7,1821	4,6215	[-16,5322, 1,7494]
PPER_PROP*	-8,7692	5,0821	[-19,1190, 0,8560]
KOUS_PROP	-18,1841	6,2338	[-31,1128, -6,4808]
BOOST_STOP_R	6,2703	2,0822	[2,2889, 10,4684]
PRED_CLASS_N	0,2655	0,0563	[0,1558, 0,3770]
BOOSTM_STOPM_R	10,4422	1,9728	[6,6104, 14,3618]
TERM_N	1,5336	0,1599	[1,2289, 1,8564]
TERM_PA1*[127]	-0,5980	0,3521	[-1,2900, 0,0930]
TERM_PO1*	0,4992	0,3160	[-0,1106, 1,1316]
TERM_GA1	-0,7606	0,3304	[-1,4131, -0,1159]
TERM_DO1*	1,0180	0,5905	[-0,0808, 2,2581]
TERM_INI	2,4236	0,2753	[1,9003, 2,9822]

Tabelle 63: Regressionskoeffizienten, Standardfehler und Konfidenzintervalle für die Modellparameter des AIC-Modells

Die Übersicht bestätigt im Wesentlichen die Ergebnisse der bivariaten Analysen, und zwar dahingehend, dass Adjektive, Partizipien (das Partizip II), Boostwörter (bzw. deren Überwiegen im Satz gegenüber eventuell auftretenden Stopwörtern), generische Prädikate (d. h. solche mit vielen GermaNet-Klassen), bestimmte morphologische Merkmale sowie Termini in bestimmten (syntaktischen) Positionen anhand ihrer Regressionskoeffizienten als positive Merkmale für das Vorliegen eines KRCs herausgestellt werden. Tendenziell scheinen darüber hinaus längere Sätze mit vielen Termini bevorzugt zu werden. Adverbien, bestimmte Pronomen, unterordnende Konjunktionen, aber auch einige, eher untypische Termpositionen sind dagegen negative Signale. Die mit * markierten Merkmale PIAT_PROP, PPER_PROP, TERM_PA1, TERM_PO1 und TERM_DO1 weisen allerdings Konfidenzintervalle auf, die über den Nullpunkt hinausgehen. Dies spricht gegen die Annahme eines signifikanten Beitrags der betreffenden Merkmale zum Modell, was für PIAT_PROP, TERM_PO1 und TERM_DO1 das Ergebnis des oben beschriebenen Signifikanztest bestätigt. Für das Merkmal PPER_PROP sind dagegen die Ausgaben von R widersprüchlich. So ergibt der von R ausgegebene standardmäßige Signikanztest im Unterschied zum Chi-Quadrat-Test auf Grundlage der Devianzdifferenzen für dieses Merkmal keine Signifikanz. Aufgrund der Ergebnisse werden jedoch die genannten Variablen als eher marginale Prädiktoren bzw. als Variablen angesehen, für deren endgültige Bewertung wiederum größere Datensätze vonnöten wären.

Wird nun nach dem Einfluss der einzelnen Prädiktoren auf die interessierende Klassifikationsentscheidung gefragt, bietet es sich zunächst an, die Regressionskoeffizienten – dies ist eine in der Interpretation logistischer Modelle häufige Praxis – als logarithmierte Chancen-Verhältnisse zu interpretieren (Sachs/Hedderich 2009[13]: 683 f.). Gegen eine solche Interpretation sprechen im hier diskutierten Fall neben den in Wolf und Best (2010: 831 ff.) vorgetragenen Argumenten auch die z. T. sehr großen Konfidenzintervalle in Tabelle 63, die eine direkte Interpretation der Koeffizienten als unsachgemäß erscheinen lassen. Aus diesen Gründen wurde eine andere Methode gewählt, um die Modellparameter hinsichtlich ihres Einflusses auf

[127] Der Index 1 scheint das Vorliegen des Merkmals zu bezeichnen.

Klassifikationsentscheidungen zu bewerten: 100 Sätze – 62 KRCs und 38 generische Sätze – wurden aus den Testdaten (s. nächsten Abschnitt) ausgewählt und auf diesen Daten wurden Klassifikationsexperimente durchgeführt. Zunächst wurden auf den Originaldaten für jeden Satz Wahrscheinlichkeiten für die Zuordnung der individuellen Beobachtungen zur Klasse der wissenshaltigen Kontexte berechnet (*setting0*), alsdann wurden – getrennt für jedes einzelne Merkmal – die Merkmalsbelegungen der Daten systematisch variiert:

- In *setting1* wurde der Anteilswert für ein gegebenes Merkmal neu berechnet, indem angenommen wurde, dass sich der Anteil beispielsweise der unterordnenden Konjuktionen (Merkmal KOUS_PROP) um 1 Konjunktion verringert (maximals bis auf 0); für binäre Merkmale wurden alle Merkmalsbelegungen auf $0 = negativ$ gesetzt, absolute Häufigkeiten wurden um den Wert 1 verringert.
- In *setting2* wurde die inverse Veränderung durchgeführt, d. h. Merkmalsbelegungen wurden auf den Wert $1 = positiv$ gesetzt oder um eine Einheit erhöht.

Auf diesen transformierten Daten wurden ebenfalls Klassifikationen errechnet. Auf den so für jedes Merkmal erarbeiteten Wahrscheinlichkeitsvektoren (für die *settings* 0-2) wurden alsdann paarweise statistische Hypothesentests (der Mann-Whitney-U-Test, vgl. Sachs/Hedderich 2009[13]: 451 ff., vgl. Abschnitt 5.2.3.4) durchgeführt, um zu überprüfen, für welche der Paare statistisch signifikante Verteilungsunterschiede, d. h. deutlich abweichende Klassifizierungen beobachtet werden können. Diese Tests ergaben nur für die Merkmale VVPP_PROP, KOUS_PROP, BOOSTM_STOPM_R, TERM_N, TERM_DO und TERM_INI signifikante Verteilungsunterschiede. Tabelle 64 zeigt diese Daten im Überblick.

Merkmal	**Durchschnittswert bzw. Verteilung der dichotomen Kategorien**	**Mittelwert Vorhersage**	**Signifikanz**
VVPP_PROP			
setting0	0,0210	0,5619	für setting1 vs. setting2
setting1	0,0043	0,5324	
setting2	0,0827	0,6372	
KOUS_PROP			
setting0	0,0030	0,5619	für setting1 vs. setting2
setting1	0	0,5710	
setting2	0,0643	0,4706	
BOOSTM_STOPM_R			
setting0	0,0710	0,5619	für setting1 vs. setting2
setting1	0,0055	0,4960	
setting2	0,1264	0,6102	
TERM_N			
setting0	1,3600	0,5619	für alle drei *settings*
setting1	0,7500	0,4504	
setting2	2,3600	0,6953	
TERM_INI			
setting0	0: 69, 1: 31	0,5619	für setting0 vs. setting2, für setting1 vs. setting2
setting1	0: 100, 1: 0	0,4916	
setting2	0: 0, 1: 100	0,7346	

TERM_DO[128]			
setting0	0: 94, 1: 6	0,5619	für setting1 vs. setting2
setting1	0: 100, 1: 0	0,5581	
setting2	0: 0, 1: 100	0,6512	

Tabelle 64: Ergebnisse der statistischen Merkmalsbewertung

Die Tabelle gibt für jedes Merkmal und *setting* einen Mittelwert für das beobachtete Merkmal bzw. – für die binären Merkmale – die Anzahlen für die jeweiligen Belegungen an. Die Mittelwerte der vorhergesagten Wahrscheinlichkeiten $\pi(x) = P(Y = 1|X = x)$ sind für jedes Experiment ebenfalls angegeben. Ganz rechts werden die Ergebnisse der Signifikanztests zusammengefasst.
Die Analyse zeigt, dass in der Regel negative Veränderungen der untersuchten Merkmale (*setting1*) im Vergleich zu *setting0* nicht zu signifikant unterschiedlichen Klassifikationen führen, was wohl auf die in diesem *setting* eher geringen Merkmalsveränderungen zurückführen ist. In der Regel führen erst Variationen um zwei Einheiten (setting1 vs. setting2), d. h. drastische Variationen zu signifikant anderen Klassifikationen. Lediglich für das Merkmal TERM_N ergeben sich signifikante Veränderungen schon bei Hinzufügen oder Entfernen eines Terminus‘, wohingegen auch für die anderen terminologischen Merkmale wiederum nur drastische Variationen zu deutlich anderen Klassifikationsergebnissen führen. Im Ergebnis zeigt sich, dass neben den lexikalischen und morphologischen Merkmalen mit den größten Regressionskoeffizienten insbesondere die terminologische Information maßgeblich für die Klassifikation eines Satzes als generisch bzw. als KRC ist, diese Merkmale können mithin als die für das hier errechnete Modell und die hier untersuchten Daten wichtigsten Variablen bezeichnet werden. Die Ergebnisse sprechen für einen termorientierten Ansatz der KRC-Extraktion, da Informationen über das Auftreten von Termini die Identifikation von KRCs zu erleichtern scheinen.

5.3.4.4 Einfaches Klassifikationsmodell

Nicht zuletzt kann das erlernte Regressionsmodel auch zum Zweck der Klassifikation ungesehenen Materials verwendet werden. Zu diesem Zweck soll ein aus Teilen des Sub-EKD und des Sub-TKD gebildetes Testset verwendet werden, das 62 KRCs und 2100 generische Sätze (der KRC-Anteil von etwa 3% entspricht ungefähr dem Anteil der KRCs in den beiden Goldstandard-Korpora) enthält. Für jeden Satz innerhalb dieser Testdaten wird dabei eine Wahrscheinlichkeit berechnet, dass es sich bei dem Satz um einen KRC handelt. Anhand dieser Wahrscheinlichkeiten können die Sätze alsdann sortiert werden. Abbildung 26 zeigt die Präzisions-Abdeckungs-Kurve, die sich bei der Auswertung dieses einfachen Rankings ergibt.
Die Kurve stellt auf der x-Achse die erreichten Abdeckungsniveaus dar, während auf der y-Achse die auf diesen Niveaus erreichte Präzision abgetragen wird. Die rote Linie visualisiert eine theoretische *baseline*, die durch zufälliges Ziehen (und Zurücklegen) von Sätzen aus dem Datensatz erreicht werden kann und bei einer durchgängigen Präzision von ca. 3% liegt. Die Kurve „LogReg“ visualisiert das Ergebnis des Rankings nach den von dem Regressionsalgorithmus ausgegebenen Wahrscheinlichkeiten. Dieses Ergebnis wird zwei weiteren, praktisch im Vergleich zu der

[128] Die sich aus der hier durchgeführten Analyse ergebende Bedeutung dieses Merkmals ist mit Vorsicht zu betrachten. Die Analyse der Trainingsdaten zeigt, dass das Merkmal in insgesamt nur 53 Fällen realisiert ist, von denen allerdings 47 KRCs sind. Umgekehrt bedeutet dies aber auch, dass alle anderen KRCs in den Trainingsdaten nicht über einen Terminus in dieser Position verfügen. Hier ist ein weiteres Mal die Bedeutung umfassenderer Untersuchungen hervorzuheben.

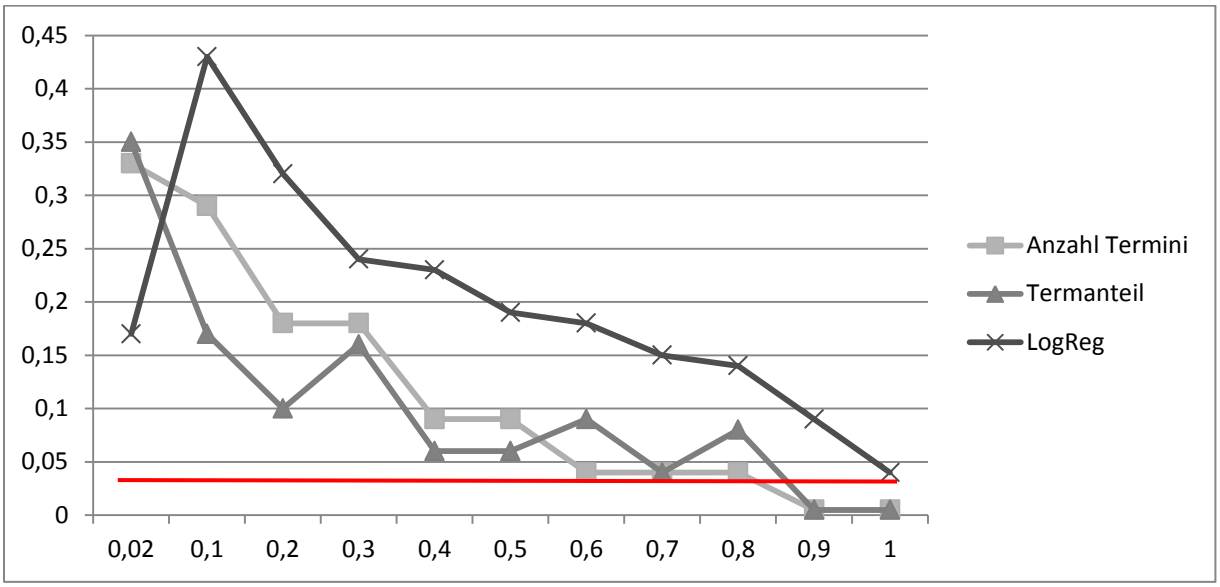


Abbildung 26: Präzisions-Abdeckungs-Kurve für ein einfaches Ranking deutscher KRC-Kandidaten aus dem Sub-EKD und dem Sub-TKD

theoretischen *baseline* eher relevanten *baselines* gegenübergestellt: „Termanteil" zeigt ein Ranking nach dem Anteil der Termini pro Satz. „Anzahl Termini" dagegen zeigt ein Ranking nach der absoluten Anzahl der Termini pro Satz[129].

Die Abbildung macht deutlich, dass nach einer Spitzengruppe mit einem Trefferanteil von über 40% die Präzision kontinuierlich abfällt und bereits bei einer Abdeckung von ca. 50% unter 20% fällt. Danach sinkt die Präzision weiter, bis sie – bei dann allerdings vollständiger Abdeckung – ein zufälliges Niveau (also das der theoretischen *baseline*) erreicht. Deutlich wird dabei, dass eine solche Klassifikation für die praktische Extraktion von KRCs aus Textkorpora nicht ausreichend ist. Allerdings war die Erarbeitung eines Extraktionsalgorithmus' auch nicht das Ziel des hier beschriebenen Experiments, vielmehr sollte überprüft werden, ob anhand rein linguistischer Merkmale (die vielfach untersuchten musterbezogenen Informationen wurden hier bewusst ausgespart) bereits eine Trennung zwischen den Klassen KRC vs. generisch beschrieben werden kann, ob also die wissenshaltigen Kontexten zugeschriebenen und durch die in diesem Kapitel durchgeführten Analysen bestätigten bzw. aufgezeigten Merkmale tatsächlich die Charakterisierung wissenshaltiger Kontexte als einen distinkten Aussagentypus rechtfertigen, und zwar unabhängig von der Art des untersuchten Korpus'. Aus diesem Grund erfolgt auch eine Untersuchung aggregierten Materials aus dem Sub-EKD und dem Sub-TKD gegenüber generischen Sätzen, wohingegen in praktischen Extraktionskontexten die Unterscheidung Entwicklungs- vs. Testkorpus, d. h. der Übergang von bekanntem auf ungesehenes Material eine größere Rolle spielen wird (vgl. hierzu die Experimente in Kapitel 6). Wird nun also bedacht, dass 25 (40%) der KRCs in dem untersuchten Datensatz eine Wahrscheinlichkeit $\pi(x)$ mit $\pi(x) = P(Y = 1|X = x) \geq 0{,}9485$ zugeordnet wurde, die

[129] Die beiden termbezogenen *baselines* sind insofern als Approximationen zu betrachten, als beide Rankings einer gewissen Willkür unterliegen, da in beiden Fällen Beobachtungen identische Werte aufweisen können, nach denen alsdann einer bestimmten Strategie gemäß zu sortieren ist. Für diese *baselines* werden daher aggregierte Präzisionswerte für Gruppen von KRC-Kandidaten angegegeben, d. h. für eine Gruppe von KRC-Kandidaten mit identischen Ranking-Koeffizienten auf dem entsprechendem Abdeckungsniveau wird der Anteil der Treffer pro Gruppe angegeben. Darüber hinaus beträgt die Präzision für das Ranking nach dem Anteil der Termini pro Satz auf dem ersten Abdeckungsniveau (0,02) p=1, da in dieser Gruppe lediglich ein valider KRC enthalten ist. Um eine übersichtlichere Gesamtdarstellung zu erhalten, wurde dieser Wert in der Darstellung allerdings nach unten korrigiert.

verbleibenden 2052 Sätze (95%) des Testdatensatzes dagegen unterhalb dieses Werts gerankt wurden, dann scheint eine Trennung der beiden Klassen tatsächlich vorzuliegen: Wird der Wert $\pi(x) = 0{,}9485$ – relativ arbiträr – als Schwellenwert für die Klassifikation eines Satzes als KRC angesetzt, ergibt sich mithin eine Klassifikationsgenauigkeit (*accuracy a*) gemäß folgender Berechnung:

$$a = \frac{Anzahl\ der\ korrekt\ klassifizierten\ Sätze}{Anzahl\ aller\ klassifizierten\ Sätze}$$

$$= \frac{25\ KRCs + 2052\ niedriger\ gerankte\ Sätze - 37\ falsch\ klassifizierte\ KRCs}{2162} = \frac{2040}{2162} \approx 94\%$$

Werden diese Ergebnisse dagegen auf die Evaluationsmaße Präzision und Abdeckung bezogen, ergibt sich eine Präzision von 23% und eine Abdeckung von 40%, mithin ein f-Maß von 0,29, so dass dieses Ergebnis im Bereich des besten Ergebnisses von Cramer (2011, vgl. Abschnitt 3.4.2.1) liegt. Anzumerken ist jedoch, dass die Ergebnisse nicht unmittelbar miteinander vergleichbar sind (insbesondere auch deshalb, weil der Anteil von Definitionen in Cramers Datensatz noch weitaus geringer ist als in den hier untersuchten Datensätzen) und die Erarbeitung eines optimalen Klassifikationssystems – hierauf wurde bereits hingewiesen – auch nicht das Ziel der hier beschriebenen Analyse darstellte. Eine – allerdings nur kursorische – Analyse falsch klassifizierter Sätze legt dabei nahe, dass zu den Ursachen für allzu niedrig gerankte KRCs Termerkennungsfehler zählen, wohingegen fälschlicherweise sehr positiv bewertete generische Sätze sich tendenziell durch eine hohe Satzlänge und Terminusdichte auszeichnen. Da es sich bei den hier untersuchten Daten um *web*-Korpora handelt, können solche Phänomene insbesondere bei Listen und ähnlichen Überblickstexten zutage treten, so dass eine Verbesserung der Vorverarbeitung möglicherweise bereits bessere Ergebnisse zeitigen kann.

5.3.4.5 Zusammenfassung

Die in diesem Abschnitt beschriebene multivariate Analyse hat deutlich gemacht, dass wissenshaltige Kontexte – jenseits von *knowledge patterns* – durch relativ einfache Merkmalskonfigurationen beschrieben werden können. Als wichtigste Merkmale haben sich dabei grammatische – Passivgebrauch und Abwesenheit von Schachtelsätzen, gemessen an den entsprechenden lexikalischen Signalen – sowie morphologische und terminologische Informationen erwiesen – die in der Literatur vielbeachteten lexikalischen Merkmale – auf deren geringe Effektstärken bereits im Zusammenhang mit den bivariaten Analysen hingewiesen worden war – scheinen dagegen aufgrund ihres insgesamt eher seltenen Auftretens von untergeordneter Bedeutung zu sein: Sie verändern erst bei starken Variationen die Chance eines Satzes, als KRC klassifiziert zu werden, deutlich. Weitere Merkmale wie etwa Generizität (Zuordnung vieler unterschiedlicher GermaNet-Klassen zu einem Verb) oder das Vorliegen vieler Nominalphrasen (hoher Anteil von Adjektiven) tragen ebenfalls zur Klassifikation bei. Zusätzlich zu dieser globalen Bewertung der Ergebnisse ist zudem von Interesse, dass die für die terminologischen Merkmale ermittelten Resultate nicht nur für die Anwendung eines termorientierten Ansatzes sprechen. Die negative Bewertung des Merkmals TERM_GA und die uneindeutige Einordnung einiger anderer terminologischer Merkmale scheinen auch darauf hinzudeuten, dass neben der reinen Anzahl der in einem KRC-Kandidaten vorgefunden Termini insbesondere die satzinitiale (Subjekt-)Position und hiermit wiederum ein linguistisch unmarkiertes Merkmal für eine positive Klassifikationsentscheidung ausschlaggebend ist. Auch dieser Befund

unterstreicht folglich die bereits im Ergebnis der uni- und bivariaten Analysen hervorgehobene linguistische Unauffälligkeit wissenshaltiger Kontexte.
Ersichtlich wird aus der durchgeführten Analyse demzufolge, dass zwar eine gewisse Separation der Klassen KRC vs. generischer Satz aufgrund der identifizierten linguistischen Merkmale möglich ist, allerdings starke linguistische Überschneidungen zwischen beiden existieren. Eine mögliche Anwendungsperspektive für ein einfaches Ranking-Verfahren wie das hier demonstrierte liegt folglich weniger in konkreten Extraktionsapplikationen, sondern vielmehr in der Vorsortierung von Korpusdaten im Vorfeld präziserer Extraktionsalgorithmen, d. h. ein Ranking oder eine Klassifikation würde dann nicht *nach* der Extraktion von KRC-Kandidaten zur Bereinigung der Ergebnisliste durchgeführt werden, sondern *vor* der Extraktion zur Verkleinerung des Suchraums zum Einsatz kommen. Nach Kenntnis der Autorin wurden ähnliche Experimente noch nicht durchgeführt. Ob eine solche Vorgehensweise tatsächlich zu besseren Ergebnissen führt, ist allerdings noch zu überprüfen. Zu beachten ist des Weiteren, dass die hier beschriebene logistische Modellierung nach wie vor eine idealisierte Beschreibung wissenshaltiger Kontexte liefert, die Raum für die Erarbeitung gesicherter Aussagen auf Grundlagen wiederholter Experimente auf größeren Datensätzen lässt, in deren Rahmen beispielsweise auch Interaktionen zwischen den identifizierten linguistischen Merkmalen oder der Einfluss möglicher Moderatorvariablen auf die Modellanpassung insbesondere an den relativ stark abweichenden Rändern des Modells untersucht werden könnten. Insgesamt kann aber festgehalten werden, dass die multivariate Analyse von KRCs mit Hilfe eines Regressionsverfahrens Einblicke in die innere Struktur der untersuchten Daten und insbesondere Hinweise auf die Eignung der postulierten *features* für praktische Extraktionsaufgaben erbracht hat.

5.4 Fazit

In diesem Kapitel wurden theoriebasiert Hypothesen bzw. Fragen zu linguistischen Eigenschaften wissenshaltiger Kontexte formuliert und alsdann empirisch mittels klassischer korpuslinguistischer sowie qualitativer Methoden überprüft. Im Rahmen dieser Untersuchungen konnten die Aussagen der analysierten themenspezifischen Literatur in vielerlei Hinsicht – insbesondere in Bezug auf lexikalische Merkmale – bestätigt, aber auch – insbesondere im Hinblick auf grammatische und strukturelle Merkmale – differenziert und ergänzt werden. Der Vergleich russischer und deutscher Daten deutet dabei auf gewisse sprachübergreifende Ähnlichkeiten hin. Die Untersuchung von Musterbestandteilen für *knowledge patterns* ergab zudem sowohl für das Russische als auch das Deutsche Übereinstimmungen mit in der Literatur beschriebenen Musterkatalogen, wobei allerdings gezeigt wurde, dass in Anbetracht der durchaus nicht immer spezifischen Verteilung dieser Einheiten auf KRCs und generische Sätze Zweifel an deren diskriminatorischer Kraft, d. h. der Nützlichkeit zu Extraktionszwecken, angebracht sind. Des Weiteren wurde ein statistisches Modell wissenshaltiger Kontexte im Vergleich zu generischen Aussagen errechnet, mit dessen Hilfe die ermittelten *features* in Bezug auf ihre Tauglichkeit für die Klassifikation von Kontext-Kandidaten bewertet wurden. Im Ergebnis erscheinen wissenshaltige Kontexte aus linguistischer Sicht als eher unauffällige Einheiten, die zwar anhand bestimmter Tendenzen und auch Abwesenheiten beschrieben, jedoch nicht vollständig von generischen Aussagen abgegrenzt werden können. Terminologische und strukturell-grammatische Merkmale scheinen dabei für KRCs von größerer Bedeutung zu sein als lexikalische Eigenschaften, wobei deutlich wird, dass eine Tendenz zur Verwendung regulärer, d. h stilistisch unmarkierter sprachlicher Mittel besteht, so dass Aussagen aus der Literatur tendenziell bestätigt werden können. Im Ergebnis scheinen für die Identifikation wissenshaltiger Kontexte tatsächlich

nicht nur linguistische, sondern auch außersprachliche Faktoren wie die von Bierwisch und Kiefer (1969) sowie Cramer (2011) genannten eine Rolle zu spielen. Die Interaktion zwischen diesen Aspekten und den linguistischen Merkmalen von Kontextkandidaten ergibt sich mithin als weiteres offenes Forschungsthema. Für eine Charakterisierung wissenshaltiger Kontexte als eigener Aussagentypus konnten zwar sprachübergreifende Hinweise gefunden werden, allerdings reichen diese für eine eindeutige Abgrenzung von generischen Aussagen nicht aus.

6. Extraktion wissenshaltiger Kontexte

6.1 Einführung

In diesem Kapitel sollen Fragestellungen in Bezug auf die Extraktion wissenshaltiger Kontexte aus Textkorpora untersucht werden. Aus den bisherigen Ausführungen ergeben sich dabei die folgenden konkreten Fragen:

- Ergeben sich bei der Extraktion russischer wissenshaltiger Kontexte besondere Schwierigkeiten, beispielsweise aufgrund der reichen Flexionsparadigmen des Russischen und seiner im Vergleich zum Deutschen noch freieren Wortfolge? Welche Performanz kann mit einem musterbasierten Extraktionsansatz erreicht werden?
- Können die im letzten Kapitel ermittelten linguistischen Merkmale wissenshaltiger Kontexte genutzt werden, um die Performanz von Extraktionssystemen zu verbessern?
- Wie nützlich sind automatisch extrahierte KRCs aus terminologischer Sicht?

Diese Aspekte sollen in den nächsten Abschnitten empirisch untersucht werden. Dabei werden zunächst sowohl für das Deutsche als auch das Russische einfache musterbasierte Systeme implementiert, mit deren Hilfe KRC-Kandidaten extrahiert werden können. Auf Grundlage der Evaluation (einschließlich einer Fehleranalyse) der dabei erhaltenen Ergebnisse sollen alsdann Ansätze entwickelt werden, mit deren Hilfe musterbasierte Verfahren in ihrer Leistungsfähigkeit gesteigert werden können. Hierbei sollen insbesondere auch die linguistischen Merkmale von KRCs im Zusammenhang mit Verfahren des maschinellen Lernens in den Blick kommen. Nicht zuletzt soll eine kritische qualitative Analyse von Extraktionsergebnissen im Hinblick auf ihren Nutzen für die praktische Terminologiearbeit dabei helfen einzuschätzen, in welchem Maße korpusbasierte Verfahren für die Terminographie genutzt werden, welche Erkenntnisse sie liefern können bzw. in welchen Punkten sie noch weiter verbessert werden müssen.

In Anbetracht der in Kapitel 3 referierten, zahlreichen computerlinguistischen Ansätze zur Extraktion wissenshaltiger Kontexte ist indes zu begründen, warum anstelle eines ML-Ansatzes anfänglich ein musterbasiertes Verfahren zur Extraktion von KRCs gewählt wird. In diesem Zusammenhang ist zunächst auf die zahlreichen hybriden und auch ML-Ansätze zu verweisen, die nach wie vor musterbezogene Informationen nutzen, so dass von einer Obsoleszenz dieser Informationen keine Rede sein kann. Des Weiteren wird in den auf die musterbasierte Extraktion folgenden Untersuchungsschritten durch die Kombination musterbasierter Verfahren mit ML-Techniken der aktuelle Forschungsstand durchaus berücksichtigt. Nicht zuletzt spricht für die Verwendung eines musterbasierten Ansatzes auch ein linguistischer Grund, und zwar derjenige, dass die referierten reinen ML-Verfahren im Ergebnis meist recht einfache Modelle definitorischer Aussagen zu liefern scheinen. Beispielsweise war in Kapitel 3 dargestellt worden, dass Borgs (2009) elaborierte Methode im Wesentlichen Varianten der Regel „Nomen – sein – Artikel – Nomen“ erlernt. Die im letzten Kapitel durchgeführte Analyse hat allerdings ergeben, dass wissenshaltige Kontexte deutlich vielfältigere Strukturen aufweisen können, die mit Hilfe linguistischer Muster gut abzubilden sind. Die Frage, ob musterbasierte Verfahren der empirisch beobachtbaren Vielfalt wissenshaltiger Kontexte eher gerecht werden können als solche Ansätze, die einen großen Teil der tatsächlich vorhandenen Variation *per definitionem* aus der Untersuchung ausschließen, ist mithin legitim, und zwar insbesondere im Hinblick darauf, dass für terminographische Zwecke eine möglichst gute Abdeckung

der Extraktionsergebnisse von großer Bedeutung ist. Des Weiteren hat bereits der in Kapitel 3 geleistete Literaturüberblick gezeigt, dass die in entsprechenden Veröffentlichungen häufig vorgetragene Behauptung einer Überlegenheit von ML-Ansätzen gegenüber linguistisch motivierten Verfahren in Anbetracht der publizierten Extraktionsergebnisse nicht immer gerechtfertigt scheint.

6.2 Musterbasierte Extraktion wissenshaltiger Kontexte

6.2.1 Implementierung der Extraktionsmuster

In dieser Arbeit wird ein syntaxbasierter Ansatz der Formulierung von Extraktionsmustern verfolgt. Für eine solche Vorgehensweise spricht nicht nur der Umstand, dass Walter (2010) bereits für das Deutsche zeigen konnte, dass der Einbezug syntaktischer Information für die von ihm untersuchten deutschen Daten zu besseren Ergebnissen führt, sondern auch die Hypothese, dass russische Strukturen aufgrund der freien Wortfolge des Russischen mittels syntaktischer Informationen effektiver zu beschreiben sind. Inwieweit dies trotz des im letzten Kapitel vorgetragenen Befundes, dass russische KRCs „reguläre“ Strukturen mit satzeröffnendem Definiendum in Subjektposition bevorzugen, möglicherweise zutrifft oder nicht, ist dabei Teil der zu untersuchenden Fragestellung.

In welcher Form können syntaxbasierte Extraktionsmuster nun implementiert werden? Ebenso wie in der Arbeit Walters (ebd.) sollen hier insbesondere Valenzinformationen zur Formulierung von Extraktionsmustern genutzt werden, wobei als verbale „Ansatzpunkte“ für die Beschreibung entsprechender Muster die im vorigen Kapitel identifizierten Bausteine für *knowledge patterns* genutzt werden. Dabei kommen selbstverständlich auch – soweit in den vorhergehenden Analysen identifiziert – nicht-verbale Musterkandidaten in den Blick, die möglicherweise eher mit Hilfe sequentieller Information zu Extraktionsmustern kombiniert werden können.
Die Muster werden für die hier durchzuführende Untersuchung in Form von XSLT-Regeln implementiert. XSLT erlaubt den direkten Zugriff auf die in XML kodierten Annotationen der Goldstandard-Korpora (vgl. Kapitel 4) in einer strukturierten und effizienten Art und Weise. Dies soll nun an je einem deutschen und russischen Muster demonstriert werden. Beispiel 97) ist ein KRC aus dem Sub-EKD.

97) Das Monopile-Konzept bezeichnet die Gründung mit einem einzelnen Stahlrohr, das über eine Rammung oder Bohrung in den Baugrund eingebracht wird.

Gemäß dem beschriebenen Ansatz kann ein solcher KRC mit Hilfe eines Musters auf Grundlage des Verbs „bezeichnen“ samt Valenzrahmen beschrieben werden. Ein solches Muster in XSLT-Kodierung ist in Abbildung 27 dargestellt.
Diese Regel durchsucht Korpussätze nach dem Lemma „bezeichnen“ und speichert für jeden gefundenen Satz die Identifikationsnummer und syntaktische Dependenz des Verbs in den Variablen „verb_id“ und „verb_dep“. Wenn das entsprechende Verb seinem morphologischen *tag* zufolge im Tempus Präsens steht („[@morph[contains(., 'tense=pres')]]“) und der syntaktische Kopf eines Terminus‘ ist – der gewählte Ansatz ist also ein termorientierter, der zu dem in Abschnitt 4.2.2 thematisierten übersetzungsbezogenen Szenario der KRC-Annotation und –Extraktion und auch den Ergebnissen der im letzten Kapitel durchgeführten linguistischen Analyse passt –, dessen syntaktische Dependenz vorab in der Variable „dep“ gespeichert wurde („$dep=$verb_id“), wird der Satz als KRC extrahiert, d. h. in eine Ausgabedatei geschrieben. Neben dem KRC-Kandidaten werden dabei auch

die Identifikationsnummer des Satzes („sent_id“) und ein Code für das Extraktionsmuster („bezeichnen1“) ausgegeben.

```
<xsl:choose>
<xsl:when test="token[@lemma='bezeichnen']">
<xsl:variable name="verb_id" select="token[@lemma='bezeichnen']/@id"/>
<xsl:variable name="verb_dep" select="token[@lemma='bezeichnen']/@dep"/>
<xsl:choose>
<xsl:when test="$dep=$verb_id and token[@lemma='bezeichnen'][@morph[contains(.,
'tense=pres')]]">
<xsl:value-of select="@sent_id"/><xsl:text>&#09;bezeichnen1&#09;</xsl:text>
<xsl:value-of select="token"/><xsl:text>&#10;</xsl:text>
</xsl:when>
</xsl:choose>
</xsl:when>
</xsl:choose>
```

Abbildung 27: XSLT-Regel für die Extraktion von Beispiel 97)

Das Beispiel macht deutlich, dass im Rahmen des hier verfolgten Ansatzes nicht lediglich Sequenzen lexikalischer Marker, sondern als für KRCs aufgrund ihrer verbalen syntaktischen Köpfe als charakteristisch betrachtete Fragmente der in Kapitel 4 annotierten und in XML formatierten Syntaxbäume gesucht werden. Welche Flexibilität mit einer solchen Formalisierung von Extraktionsmustern erreicht wird, machen die folgenden Beispiele deutlich, die ebenfalls durch das in Abbildung 27 dargestellte Muster extrahiert werden.

98) Moderne Vertikal-Achser bezeichnet man unter anderem auch als H-Rotor, weil die Blätter senkrecht zur Achse oder Rotorwelle stehen.

99) Derjenige Anteil einer Wärmemenge Q, der prinzipiell in Arbeit umwandelbar ist (also hQ) wird als der Exergieinhalt von Q bezeichnet (Exergie).

Ein ähnlich simples russisches Extraktionsmuster ist in Abbildung 28 dargestellt.

```
<xsl:if test="token[@lemma[contains(., 'предназначен')]][@id=$dep]">
<xsl:value-of select="@sent_id"/><xsl:text>&#09;предназначен&#09;</xsl:text>
<xsl:value-of select="token"/><xsl:text>&#10;</xsl:text>
</xsl:if>
```

Abbildung 28: Einfache russische Extraktionsregel

Diese Regel sucht nach einem russischen Lemma, das die Zeichenfolge *предназначен* enthält (womit folglich auch die Formen *предназначена*, *предназначены* und *предназначено* sowie Formen der adjektivischen Ableitung *предназначенный* gefunden werden). Auch hier wird jedoch wieder die zusätzliche Bedingung eingeführt, dass das betreffende Lemma der syntaktische Kopf eines

Terminus‘ sein muss („@id=$dep“). Die folgenden Beispiele demonstrieren, welche KRCs unter anderem durch dieses einfache Muster extrahiert werden.

100) Сцепление предназначено для кратковременного разъединения двигателя от трансмиссии и плавного их соединения в тех случаях, когда требуется затормозить автомобиль, начать движение или переключить передачи.
[*Übersetzung:* Die Kupplung hat die Aufgabe, den Motor kurzzeitig vom Getriebe abzukoppeln bzw. störungsfrei mit diesem zu verbinden, wenn der Wagen angehalten oder in Bewegung gesetzt bzw. wenn ein neuer Gang eingelegt werden soll.]

101) Рулевое управление предназначено для обеспечения движения автомобиля по заданному водителем направлению.
[*Übersetzung*: Die Lenkung dient dazu, das Fahrzeug in die vom Fahrer gewünschte Richtung zu steuern.]

102) Тормозная система предназначена для уменьшения скорости движения и/или остановки транспортного средства.
[*Übersetzung:* Das Bremssystem dient zur Verringerung der Geschwindigkeit und/oder zum vollständigen Anhalten des Fahrzeugs.]

103) Ручной тормоз ("ручник") предназначен для того, чтобы зафиксировать автомобиль на месте при длительной стоянке, на уклоне.
[*Übersetzung:* Die Handbremse dient dazu, das Fahrzeug bei längerem Stand oder an einem Abhang zu fixieren.]

Bei der Formulierung von Extraktionsmustern für deutsche und russische KRCs wird wie folgt vorgegangen: Zunächst werden für die im vorigen Kapitel identifizierten potentiellen Musterbausteine einfache Suchanfragen formuliert. Als Datengrundlage für die konkrete Form der Anfragen werden dabei die Entwicklungskorpora Sub-EKD und Sub-EKR herangezogen. Auch redundante Muster werden in diesem Schritt in der Hoffnung berücksichtigt, mit diesen auf den Testkorpora eine bessere Abdeckung zu erreichen. Die entwickelten Anfragen können im Entwicklungsprozess iterativ auf die Entwicklungskorpora angewandt werden, wobei solche Muster, die allzu viele ungültige Ergebnisse erbringen (also ausschließlich oder überwiegend generische Sätze bei schlechter KRC-Abdeckung extrahieren), aus der Menge der Suchmuster entfernt werden können. Ein Vorteil dieser Vorgehensweise besteht darin, dass die Mustermenge bereits in dieser Phase – zumindest in Ansätzen – optimiert werden kann, wobei eine allzu einseitige Konzentration auf die Präzision allerdings zu vermeiden ist, da hierunter mit hoher Wahrscheinlichkeit die Abdeckung leidet. Vielmehr soll bei der Formulierung der Suchmuster versucht werden, eine möglichst hohe Abdeckung bei mindestens akzeptabler Präzision auf den Entwicklungsdaten zu erreichen. Letztlich bleibt aber aufgrund der Vielzahl der möglicherweise relevanten linguistischen Parameter und deren Ausprägungen – dies sei an dieser Stelle nicht verschwiegen – die manuelle Formulierung von Suchmustern ein im Vergleich zu den Regelgenerierungsverfahren des maschinellen Lernens (vgl. Witten et al. 2011[3]) nur schwer zu formalisierender Prozess.

6.2.2 Genutzte Muster

6.2.2.1 Deutsch

Für das Deutsche werden im Ergebnis der Musterformulierung insgesamt 82 Suchmuster verwendet, die in Anhang 10 zusammengefasst sind. Dort sind die Muster als Graphen dargestellt und mit einer kurzen Beschreibung in Textform sowie einem Beispiel versehen. In den Graphen bilden *tokens* bzw. Sequenzen von *tokens*, über denen Abfragen formuliert werden, die Knoten. Eigenschaften der *tokens*, beispielsweise grammatische oder lexikalische Einschränkungen, sind dabei als Knoteneigenschaften in Form von XSLT-Befehlen oder leicht vereinfachten XSLT-Befehlen dargestellt, wobei Attribute durch das Zeichen *@*, Variablen durch das Zeichen *$*, Termini bzw. Definienda durch den *string* „term“ und das jeweilige *knowledge pattern* durch die Zeichensequenz „[@lemma=““]“ gekennzeichnet werden. Die Namen und Werte der Attribute entsprechen dabei den in Abschnitt 4.4.3 beschriebenen Attributen der XML-Formatierung der Goldstandard-Korpora. Als Variablen werden die Kürzel *$id* (Identifikationsnummer eines *tokens*), *$dep* (Verweis auf die Identifikationsnummer des syntaktischen Kopfes eines *tokens*), *$verb_id* (Identifikationsnummer des *knowledge patterns*) und *$verb_dep* (Verweis auf die Identifikationsnummer des syntaktischen Kopfes eines *knowledge patterns*) verwendet. Die gerichteten Kanten zwischen den Knoten visualisieren die zwischen diesen bestehenden syntaktischen Beziehungen, wobei das Pfeilende stets den syntaktischen Kopf einer solchen Struktur markiert. Ungerichtete Kanten dienen dagegen der Darstellung zusätzlicher Bedingungen – etwa in der Form des Verbots bestimmter *tokens* („!token“) – oder Mengen von Bedingungen, von denen mindestens eine erfüllt sein muss (logisches ODER – dieses ist in den Graphen in der Regel durch die Angabe „OR“ markiert).
Die formulierten Extraktionsmuster für die Extraktion deutscher KRCs aus Textkorpora stützen sich auf insgesamt 49 *knowledge patterns* bzw. Strukturen (vgl. Anhang 10), die unterschiedlichen funktionalen Typen zugeordnet werden können, und zwar:

- *linguistischen Prädikaten zur Verbalisierung der in Kapitel 2 erarbeiteten semantischen Zielrelationen*: bestehen, gegeben sein (durch etwas), sein, darstellen, umwandeln, verwenden, beschreiben, anordnen, einsetzen, ausführen, erlauben, sich ergeben (aus), antreiben, befinden, bewirken, herstellen/hergestellt werden, auftreten, besitzen, dienen, gewährleisten, kennzeichnen/gekennzeichnet sein durch, benutzen, basieren, entsprechen, Funktionsverbgefüge mit „haben“, Funktionsverbgefüge mit dem Schlüsselwort „Prinzip“, Funktionsverbgefüge mit „sein“ oder „liegen“, kommen (zu etwas), ausrüsten/ausgerüstet sein, führen (zu), ausdrücken/ausgedrückt werden, nutzen,
- *linguistischen Prädikaten zum Ausdruck klassifikatorischer Information*: unterscheiden, gehören (zu), zählen (zu),
- *metalinguistischen Prädikaten des Definierens und Benennens*: bezeichnen, nennen, verstehen (unter), bedeuten, sprechen (von), heißen,
- *(nicht-)prädikativen linguistischen Mitteln des Thematisierens und Hervorhebens*: sogenannt, bei, in, handeln (um), es gibt,
- *syntaktischen Strukturen*: Appositionen, also, Relativsätze.

Damit wurden alle in Abbildung 17 (Abschnitt 5.2.2.5.2) dargestellten *knowledge patterns* – mit Ausnahme von Parenthesen und Doppelpunkten, die sich nur schwer in Muster mit akzeptabler Performanz umsetzen lassen – zur Formulierung von Extraktionsmustern genutzt. Darüber hinaus bestätigt dieser Überblick über implementierte *knowledge patterns* insofern das Ergebnis der in

Abschnitt 5.2.2.5.3 durchgeführten Analyse semantischer Klassen von Musterbausteinen, als tatsächlich ein großer Anteil von Verben genutzt wurde, die der GermaNet-Klasse „Allgemein“ zugeordnet werden können (z. B. „sein“, „darstellen“, „bestehen“ oder „kennzeichnen“). Im Vergleich zu Tabelle 4 (Abschnitt 3.2.3) wird allerdings deutlich, dass die Analyse von *knowledge patterns* im Sub-EKD eine deutlich geringere Anzahl insbesondere metalinguistischer Muster ergeben hat als frühere Studien. Auch in Bezug auf nicht-prädikative Mittel des Thematisierens und Hervorhebens ergibt die Literatur ein breiteres Variationsspektrum. Dagegen umfassen die auf Grundlage des Sub-EKD formulierten Suchmuster deutlich mehr Abfragen auf Grundlage von Prädikaten zur Verbalisierung spezifischer semantischer Relationen. Inwieweit dieser Effekt auf eine im Gegensatz zu dem hier verfolgten breiten Ansatz der Untersuchung wissenshaltiger Kontexte stehenden Fokussierung bzw. Einengung bisheriger Studien auf bestimmte Definitionstypen (insbesondere aristotelische Definitionen) zurückzuführen ist, kann ohne eine tiefergehende Untersuchung nicht beantwortet werden, als naheliegende Ursachen für die beobachteten Unterschiede erscheinen jedoch auch textsorten- und fachgebietsspezifische Effekte (z. B. im Hinblick auf in der Literatur genannte Muster wie „zu neudeutsch“ oder „erkennt man daran, dass“, die in den hier untersuchten weitestgehend technischen Texten als unwahrscheinlich gelten dürften). Einige der in der Literatur genannten zusätzlichen Muster (z. B. „besagen“) wurden zudem im Prozess der Musterformulierung auf ihre Tauglichkeit geprüft, jedoch aufgrund schlechter Ergebnisse verworfen. Im Ergebnis festgehalten werden kann mithin, dass Überschneidungen zwischen den in der Literatur beschriebenen und den hier entwickelten Mustermengen nur in geringem Maße existieren, was darauf hindeutet, dass Suchmuster entweder auf Grundlage deutlich größerer Datenmengen oder aber – ebenso wie andere Sprachverarbeitungs- und Analysewerkzeuge – im Hinblick auf konkrete Fachgebiete oder Textsorten entwickelt werden müssen.

6.2.2.2 Russisch

Für das Russische ergibt die Musterformulierung auf Grundlage des Sub-EKR weniger Extraktionsmuster als für das Deutsche, und zwar 54 (ein Überblick über diese Muster findet sich in Anhang 11, die Darstellung der Muster erfolgt analog zur Darstellung für das Deutsche), wobei noch keine Entfernung unpräziser Muster im Sinne von Abschnitt 6.2.1 stattgefunden hat. Diese Muster stützen sich auf ähnliche funktionale Typen wie die deutschen Muster und können 36 *knowledge patterns* bzw. Strukturen zugeordnet werden:

- *linguistischen Prädikaten zur Verbalisierung der in Kapitel 2 erarbeiteten semantischen Zielrelationen*: иметь, использовать, использоваться, корректировать, обеспечивать, оборудоваться, оборудовать, осуществляться, передавать, поддерживать, позволять, предназначать, представлять собой, применять, применяться, приводиться, расположить, служить, снижать, состоять, управлять, устанавливать, устанавливаться, установить, включить, воздействовать, работа заключается,
- *linguistischen Prädikaten zum Ausdruck klassifikatorischer Information:* делиться, различать,
- *metalinguistischen Prädikaten des Definierens und Benennens*: называться, называть,
- *(nicht-)prädikativen linguistischen Mitteln des Thematisierens und Hervorhebens*: Präpositionen,
- *grammatischen und syntaktischen Strukturen*: Agens im Instrumental, Relativsätze, Parenthese, Bindestrich-Strukturen.

Auch hier zeigt sich, dass die in Abbildung 16 (Abschnitt 5.2.2.5.1) identifizierten *knowledge patterns* zwar zum größten Teil in Muster umgesetzt wurden, im Vergleich zu den für andere Sprachen in der Literatur genannten Mustern aber wahrscheinlich nicht das volle Potential ausgeschöpft wird. Anders als für das Deutsche wurde zudem für das Russische aufgrund der angenommenen geringen Präzision einiger Suchmuster sofort eine Reihe von Filtern implementiert, welche die Extraktion ungültiger Sätze vermeiden sollen. Hierbei handelt es sich um formale ebenso wie um lexikalische (z. B. bestimmte Demonstrativa in satzinitialer Position) und grammatische (z. B. Verb in der ersten Person) Elemente.

6.2.3 Ergebnisse

6.2.3.1 Deutsch

6.2.3.1.1 Baseline

Mit den bisher beschriebenen deutschen Suchmustern lässt sich ein einfaches Extraktionssystem umsetzen, das auf die Entwicklungs- (Sub-EKD) und Testkorpora (Sub-TKD) angewandt werden kann. Die Musterimplementierung entspricht dabei der Darstellung in Anhang 10. Die Extraktionsergebnisse werden anhand der Goldstandard-Annotation evaluiert. Allerdings sind in diesem Zusammenhang noch einige Anmerkungen angebracht:

- Die Extraktionsergebnisse weisen einige wenige Sätze auf, die Goldstandard-KRCs stark ähneln, allerdings nicht als KRCs annotiert wurden. Diese Sätze werden nicht als Treffer gewertet. Desgleichen werden mögliche Zweifelsfälle, also solche Sätze, die u. U. auch als wissenshaltige Kontexte anerkannt werden könnten, aber nicht in den Goldstandard aufgenommen wurden, nicht als Treffer bewertet. Dies bedeutet, dass ein „weicheres“ Evaluationsszenario etwas bessere Präzisionswerte erbringen würde, die Abdeckung der Extraktion wäre in diesem Fall aber nur noch zu schätzen.
- Einige der extrahierten Sätze enthalten aufgrund der nicht perfekten Satzgrenzenerkennung im Zusammenhang mit der Korpusannotation (vgl. Kapitel 4) tatsächlich mehrere Sätze. Sofern es sich hierbei um generische Sätze, also *noise*, handelt, werden diese nicht als zusätzliche falsche Positive bewertet. Enthält ein extrahierter Satz jedoch mehrere KRCs, wirkt sich dies positiv auf die Abdeckung aus, die KRCs werden dann also einzeln bewertet.
- In den Goldstandard-Korpora wurden 339 (Sub-EKD) bzw. 287 (Sub-TKD) extrahierbare KRCs gefunden (vgl. Abschnitt 5.3.3), die Evaluation erfolgt automatisch anhand der Satz-Identifikationsnummern dieser Fundstellen. Die im Vergleich zu den in Kapitel 4 genannten abweichenden Zahlen ergeben sich dabei aus der Tatsache, dass während der Goldstandard-Annotation möglicherweise in den Korpora auftretende (und auch durch die Duplikatentfernung nicht bereinigte) Dopplungen nicht berücksichtigt wurden, um den Bewertungsaufwand möglichst gering zu halten, für die Extraktion sind diese Fundstellen aber als Treffer zu werten. Darüber hinaus können einige annotierte KRCs aufgrund fehlerhafter Satztrennungen nicht so extrahiert werden, dass sie die Funktion der Wissensvermittlung noch erfüllen können. Diese Sätze werden daher nicht zur Bewertung der Abdeckung herangezogen und im Fall der Extraktion von Fragmenten auch als falsche Positive bewertet.

- Von unterschiedlichen Mustern mehrfach extrahierte Sätze werden für die Berechnung der globalen Maße Präzision, Abdeckung und f-Maß nur einfach evaluiert. Zur Bewertung der Performanz der einzelnen Extraktionsmuster wird dagegen die Gesamtmenge der von einem Muster extrahierten Sätze herangezogen.

Die Extraktionsergebnisse auf den Korpora werden mittels der in Kapitel 3 eingeführten Evaluationsmaße Präzision, Abdeckung und f-Maß bewertet. Dabei ergeben sich die in Tabelle 65 zusammengefassten Werte.

Evaluationsmaß	**Sub-EKD**	**Sub-TKD**
Präzision	0,4212 (584 extrahierte Sätze)	0,4654 (318 extrahierte Sätze)
Abdeckung	0,7257 (246 extrahierte KRCs)	0,5157 (148 extrahierte KRCs)
f-Maß	0,5330	0,4893

Tabelle 65: Performanz des Baseline-Extraktionssystems auf den deutschen Goldstandard-Korpora

Abbildung 29 zeigt den Präzisions-Abdeckungs-Graphen, der die jeweils beste erreichte Präzision auf verschiedenen Abdeckungsniveaus für die beiden Korpora darstellt.

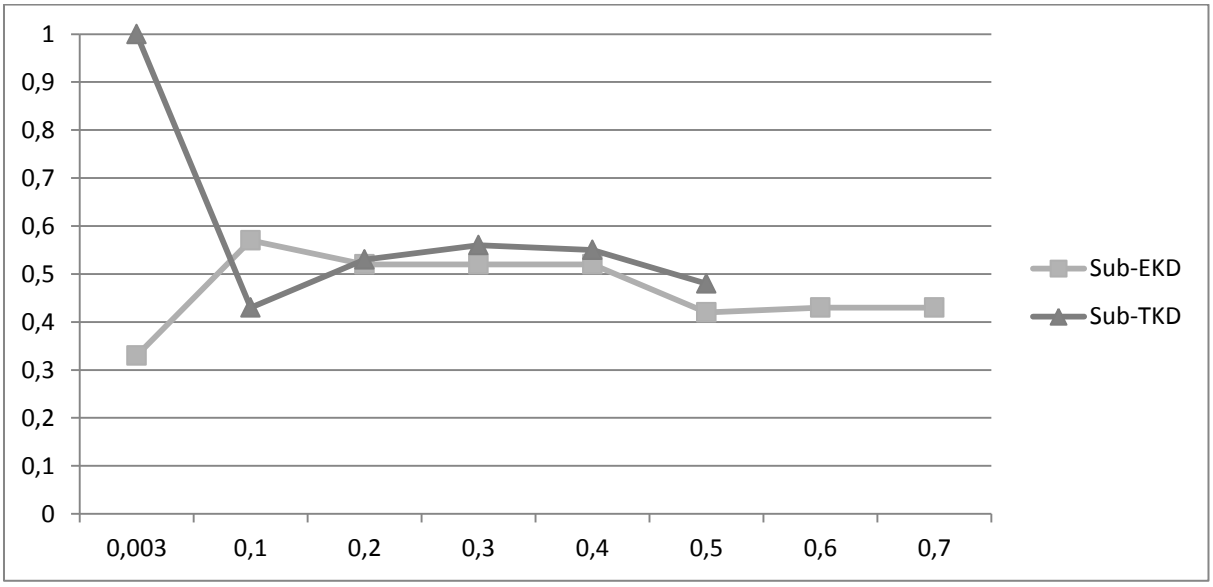


Abbildung 29: Präzisions-Abdeckungs-Graph für das deutsche Baseline-Extraktionssystem

Abbildung und Tabelle zeigen, dass für beide Korpora relativ konstante Präzisionswerte in einem mittleren Performanzbereich erreicht wurden, wobei die Präzision für das Testkorpus – bei allerdings deutlich schlechterer Abdeckung – sogar leicht über der für das Entwicklungskorpus erreichten liegt. Die Werte liegen über den in der Literatur für das Deutsche genannten (vgl. Kapitel 3), ohne jedoch mit diesen direkt vergleichbar zu sein. Die Graphik zeigt zudem, dass die Performanz des musterbasierten Extraktionsansatzes auch durch Eigenschaften der Korpora, aus denen extrahiert wird, beeinflusst wird: Die Extraktionsergebnisse werden in der Reihenfolge ihrer Satz-Identifikationsnummern bewertet und bilden damit die inhaltliche und linguistische Progression der in den Korpora zusammengetragenen Texte ab. Mithin liegt es nahe, leichte Schwankungen in der erreichten Präzision auch auf Eigenschaften der einander ablösenden Texte zurückzuführen.

Die verwendeten Extraktionsmuster können zudem in Bezug auf ihre Performanz bewertet werden. Tabelle 66 zeigt eine solche Evaluation für das Maß Präzision, das hier als Quotient aus allen durch ein Muster extrahierten Treffern (einschließlich redundanter, also auch durch andere Muster

extrahierte Treffer) und allen durch ein Muster extrahierten Sätzen (Treffer und falsche Positive) insgesamt berechnet wurde.

Muster	**Präzision Sub-EKD**	**Präzision Sub-TKD**
also	*1,00*	-
anordnen1	*0,57*	*0,00*
anordnen3	0,33	-
antreiben1	*0,67*	*0,00*
Apposition1	0,23	0,00
Apposition2	*1,00*	-
Apposition3	0,20	0,33
auftreten	*1,00*	-
ausdrücken2	*1,00*	-
ausführen1	*1,00*	*0,00*
ausführen3	*1,00*	-
ausführen4	*1,00*	-
ausrüsten	0,25	-
basieren	*0,50*	-
bedeuten	*1,00*	-
befinden	**0,62**	**0,50**
bei1	0,34	0,11
bei2	0,41	0,33
benutzen	*1,00*	*0,33*
beschreiben1	*1,00*	*0,00*
beschreiben2	**0,67**	**1,00**
besitzen	*0,14*	*0,63*
bestehen1	**0,72**	**0,93**
bestehen3	**1,00**	**1,00**
bestehen5	*0,75*	-
bestehen6	*1,00*	-
bewirken2	*1,00*	*0,00*
bezeichnen1	**1,00**	**1,00**
bezeichnen2	**1,00**	**1,00**
bezeichnen3	**0,63**	**0,90**
bezeichnen5	*1,00*	-
bezeichnen7	**1,00**	**1,00**
darstellen1	*0,50*	*0,22*
darstellen4	0,38	0,33
darstellen5	*0,67*	-
dienen1	*0,25*	*1,00*
dienen2	*1,00*	-
einsetzen1	**1,00**	**0,73**
einsetzen3	0,25	0,00
entsprechen	0,33	0,00
ergeben_sich1	*1,00*	-
ergeben_sich2	*1,00*	-
erlauben	**0,75**	**0,50**
es_gibt	0,23	0,22
etwas_sein	*0,80*	*0,00*
führen	0,29	0,25

gegeben_sein	*1,00*	-
gehören1	**0,67**	**0,50**
gewährleisten2	*1,00*	-
haben	*1,00*	*0,00*
handeln_um	**0,80**	**0,50**
heißen	*0,33*	*0,50*
herstellen	*0,60*	*0,00*
in	0,47	0,36
kennzeichnen	*1,00*	-
kommen_zu	*1,00*	-
nennen1	**0,50**	**0,50**
nennen2	**1,00**	**1,00**
nennen3	*0,43*	*1,00*
nennen4	*1,00*	-
nennen5	*1,00*	-
nennen6	*1,00*	-
nutzen4	0,22	0,33
prinzip	*0,67*	-
Relativ	0,16	0,33
sein1a	0,29	0,45
sein1b	0,42	0,40
sogenannt1	**0,50**	**0,94**
sogenannt2	**0,67**	**0,86**
sprechen	*0,33*	*1,00*
umwandeln1	*1,00*	*0,00*
umwandeln2	*1,00*	-
umwandeln4	*0,80*	*0,00*
unterscheiden1	**1,00**	**0,50**
unterscheiden2	**1,00**	**1,00**
unterscheiden4	*0,67*	*0,00*
unterscheiden5	*1,00*	-
verstehen	*1,00*	-
verwenden2	**0,57**	**1,00**
verwenden3	*1,00*	-
verwenden4	*0,67*	*0,00*
zählen	**0,67**	**0,50**
durchschnittliche Präzision	**0,72**	**0,45**

Tabelle 66: Präzision der deutschen Extraktionsmuster auf den Entwicklungs- und Testdaten

Die Tabelle zeigt zweierlei:

- Viele der auf dem Entwicklungskorpus entwickelten Muster tragen nicht zur KRC-Extraktion aus dem Testkorpus bei (entsprechend enthält die rechte Tabellenspalte für diese Muster keine Präzisionsangabe). Gleichzeitig scheinen für dieses relevante Suchmuster zu fehlen, so dass auf dem Testkorpus eine deutlich (um ca. 20 Prozentpunkte) niedrigere Abdeckung erreicht wird. Der Hauptteil derjenigen Muster, die auf den Testdaten versagen, weist zudem auf dem Entwicklungsdatensatz sehr hohe Präzisionswerte auf. Insgesamt scheint dies auf eine Überanpassung der Extraktionsmuster an die Entwicklungsdaten durch Formulierung allzu spezifischer oder aber generell (d. h. bereits in Bezug auf das *knowledge* pattern) nicht auf andere Datensätze verallgemeinerbare Muster hinzudeuten. Wahrscheinlich auf diesen

Umstand ist auch der große Unterschied in der durchschnittlichen Musterpräzision zwischen Entwicklungs- und Testkorpus (letzte Tabellenzeile) bei allerdings ähnlicher Extraktionspräzision (Tabelle 65) zurückzuführen: Zwar fallen auf dem Testkorpus relativ viele Muster weg, die auf dem Entwicklungskorpus sehr hohe Präzisionswerte erreicht haben, wodurch die durchschnittliche Musterpräzision sinkt. Allerdings war deren Beitrag ohnehin nur gering, so dass die Gesamtpräzision der Extraktion aus dem Sub-TKD nicht leidet.

- Viele der verbleibenden Muster zeigen im Vergleich von Entwicklungs- und Testdaten starke Schwankungen (> 0,5) in Bezug auf die ermittelten Präzisionswerte. Diese Muster sind in der Tabelle kursiv markiert, sie können als nicht zuverlässig angesehen werden. Die verbleibenden Muster mit einer stabilen Präzision von mindestens 50% auf beiden Datensätzen sind in der Tabelle fett markiert und nicht sehr zahlreich. Es handelt sich hierbei tatsächlich um generische Extraktionsmuster, die allerdings allein kaum eine zufriedenstellende Extraktionsperformanz garantieren können. Durchgängig niedrigperformante Muster sind unmarkiert.

Dass generische, hochperformante Extraktionsmuster allein keine ausreichende Performanz erbringen, kann experimentell bestätigt werden. Werden nur diejenigen Extraktionsmuster verwandt, die auf dem Entwicklungskorpus eine Präzision von mindestens 50% erreichen, ergibt sich zwar für das Sub-EKD ein leicht gesteigertes f-Maß von 0,57 (73% Präzision bei einer Abdeckung von 46%), jedoch sinkt das f-Maß für das Testkorpus auf 0,37 (65% Präzision bei einer Abdeckung von 26%). Werden dagegen Suchmuster nach den Kriterien Abdeckung oder f-Maß ausgewählt, ergibt sich stets eine schlechtere Gesamtperformanz. Mithin führt das Entfernen von Mustern aus der Suchmustermenge in der Regel zu schlechteren Ergebnissen. Die Daten scheinen auch darauf hinzudeuten, dass für eine mindestens akzeptable Abdeckung „schmutzige“ Muster, d. h. solche mit geringer Präzision, aber dennoch nennenswertem Beitrag zur Abdeckung (etwa Relativsätze oder die Muster auf „bei“), verwendet werden müssen. Die Optimierung musterbasierter Ansätze erscheint folglich im Vergleich zu einer Vor- oder Nachfilterung von KRC-Kandidaten als eher schwierig umsetzbar. In den folgenden Abschnitten einige Ansätze zur Verbesserung der musterbasierten Extraktionsergebnisse beschrieben werden, wobei – dies sei der Genauigkeit halber gesagt – Optimierungsverfahren im mathematischen Sinn nicht zur Anwendung kommen.

6.2.3.1.2 Punktuelle Präzisionsverbesserung

In diesem Abschnitt sollen einige Korrekturen des *baseline*-Systems beschrieben werden, mit deren Hilfe leichte Präzisionsverbesserungen erreicht werden konnten. Formalisierte und verallgemeinerbare Algorithmen, die dies leisten, werden jedoch nicht beschrieben, vielmehr beschränkt sich die Darstellung auf eine knappe Zusammenfassung derjenigen Optimierungsschritte, mit denen konsistente – wenn auch geringfügige – Verbesserungen erzielt werden konnten. Eine eher algorithmisch ausgerichtete Beschreibung von Optimierungsversuchen findet sich bei Walter (2010: 206 ff.). Walter testet sowohl ein Verfahren der inkrementellen Musterauswahl zur Bestimmung einer optimalen Suchmustermenge als auch einen lexikalischen Filter, kann allerdings nur in einem einzigen Experiment eine Verbesserung des f-Maßes (von 0,27 auf 0,3, ebd.: 209) beim Übergang auf ungesehene Daten erreichen.

Für das hier verfolgte Vorhaben einer punktuellen Präzisionsverbesserung wurden einerseits einige wenige, simple Ausnahmeregeln implementiert, bei deren Vorliegen Sätze nicht als KRC-Kandidaten

aus den Korpora extrahiert werden sollen. Diese Regeln umfassen insbesondere Einschränkungen auf Grundlage von Annahmen zur linguistischen Form wissenshaltiger Kontexte. So dürfen KRC-Kandidaten weder mit einem Demonstrativpronomen noch mit einem Pronominaladverb, einer Kardinalzahl, einem Personalpronomen (mit Ausnahme von „es“) oder einer unterordnenden Konjunktion beginnen. Weitere Regeln dienen der Identifikation falscher Positive mit formalen Defekten oder in KRCs unüblichen Lemmata wie „ich“ oder „wir“. Da hier, wie bereits ausgeführt, weniger die Entwicklung entsprechender Algorithmen im Vordergrund stand als die Definition von Regeln, die tatsächlich zu Verbesserungen – und hierin besteht eine nennenswerte Schwierigkeit entsprechender Optimierungsversuche – auf beiden Datensätzen führen, wurde die Wirksamkeit dieser Einschränkungen sowohl auf dem Sub-EKD als auch auf dem Sub-TKD automatisch evaluiert, so dass nur diejenigen Regeln implementiert wurden, mit denen tatsächlich eine Verbesserung erreicht werden konnte.

Andererseits wurde versucht, die Präzision einer Reihe unpräziser Suchmuster zu steigern bzw. unpräzise Muster aus der Suchmustermenge zu entfernen, ohne dass die Abdeckung der Extraktion allzu stark darunter leidet. Dabei wurden folgende Schritte umgesetzt:

- Modifikation von Mustern: *beschreiben1* muss nun eine zweite Nominalphrase enthalten, die sich ebenfalls auf das *knowledge pattern* bezieht; *führen* muss im Tempus Präsens stehen; im Muster *herstellen* muss das Hilfsverb sich vor dem *knowledge pattern* befinden; im Muster *sprechen* wurde die Verwendung der Adverbien „hier“ und „heute“ verboten.
- Löschen von Mustern: Gelöscht wurden die Muster *darstellen4, entsprechen, Apposition3, nutzen4*.

Offensichtlich handelt es sich bei diesen Verbesserungsschritten um sehr einfache Maßnahmen. Mit Hilfe dieser Veränderungen konnte indes die Performanz des *baseline*-Systems leicht gesteigert werden. Tabelle 67 stellt die neuen Werte den älteren Werten aus Tabelle 65 gegenüber.

Evaluationsmaß	**Sub-EKD (alt)**	**Sub-EKD (neu)**	**Sub-TKD (alt)**	**Sub-TKD (neu)**
Präzision	0,4212 (584 Sätze)	0,5011 (461 Sätze)	0,4657 (318 Sätze)	0,5422 (249 Sätze)
Abdeckung	0,7257 (246 KRCs)	0,6814 (231 KRCs)	0,5157 (148 KRCs)	0,4704 (135 KRCs)
f-Maß	0,5330	0,5775	0,4893	0,5038

Tabelle 67: Performanz des Extraktionssystems auf den deutschen Goldstandard-Korpora nach Präzisionsoptimierung

6.2.3.1.3 Möglichkeiten der punktuellen Abdeckungsverbesserung

Für eine punktuelle Verbesserung der Abdeckung des beschriebenen *baseline*-Extraktionssystems bieten sich zwei Möglichkeiten an:

- Hinzunahme neuer Extraktionsmuster,
- Modifikation oder Erweiterung bereits implementierter Extraktionsmuster mit geringer oder unvollständiger Abdeckung zur Erfassung weiterer KRCs.

Walter (ebd.: 212 ff.) versucht einerseits, manuell zusätzliche Suchmuster zu spezifizieren, und andererseits, mittels *bootstrapping* (ebd.: 235 ff.) automatisch neue Extraktionsmuster zu erlernen. Für die manuelle Musterspezifikation wählt er dabei häufige Verben nach semantischen Kriterien als Musterkandidaten aus und spezifiziert auf dieser Grundlage neue valenzbasierte Muster. Dieser Prozess ergibt 44 neue Suchmuster (ebd.: 213). Allerdings erzielt der Autor damit auf den von ihm untersuchten Daten stets eine Verschlechterung des f-Maßes (ebd.):

> „In dem von uns betrachteten konkreten Testfall haben die zusätzlich spezifizierten Suchmuster […] keine Verbesserung der Ergebnisqualität erbracht. Wie der sehr geringe Recall-Wert von 0,03 für die Extraktion mit den neuen Suchmustern alleine erkennen lässt, sind unter den zusätzlichen Definitionsverben offenbar nur wenige, die überhaupt in Definitionen in unserem Goldstandard-Korpus auftreten."

Aus diesen Ergebnissen zieht der Autor (ebd.: 215) die Schlussfolgerung,

> „[…] dass der Recall-Verlust beim Übergang vom Entwicklungs- auf das Testkorpus auch auf Auftreten seltener, mehr oder weniger idiosynkratischer Definitionsformulierungen zurückzuführen ist und somit nicht allein durch die Ermittlung weiterer häufiger Definitionsmuster behoben werden kann. Nicht zuletzt aus diesem Grund ist auf der Basis einer intellektuellen Analyse von Korpusbeobachtungen wohl generell nur ein begrenzter Grad an Vollständigkeit bei der Suche erreichbar."

Dieser Befund deckt sich insofern mit der oben vorgetragenen Analyse der Musterperformanz, als einerseits durch Muster, die auf idiosynkratischen Formulierungen im Entwicklungskorpus beruhen, keine neuen KRCs aus dem Testkorpus extrahiert werden können (also eine Überanpassung der Muster an das Entwicklungskorpus vorliegt) und andererseits z. T. idiosynkratische Formulierungen im Testkorpus die Leistungsfähigkeit selbst generischer Muster einschränken. Darüber hinaus scheinen auch gewisse fachspezifische Unterschiede eine Rolle zu spielen. Insgesamt legen die Ergebnisse wiederum eine Suche nach Extraktionsmustern in größeren Datenmengen nahe, mit deren Hilfe mehr generische, d. h. universell einsetzbare Muster gefunden werden könnten. Walter (ebd.: 235 ff.) setzt eine solche Suche mittels des bereits (Kapitel 3) beschriebenen automatischen *bootstrapping*-Ansatzes um. Der Autor (ebd.: 243) kann mit Hilfe dieses Verfahrens die Abdeckung seiner Extraktion deutlich stärker steigern als durch die manuelle Spezifikation zusätzlicher Muster, wobei allerdings auch ein Präzisionsverlust eintritt (diesen fängt Walter allerdings durch ein automatisches Ranking-Verfahren auf, mit dem er schließlich signifikant bessere Ergebnisse erzielt als mit seinem *baseline*-System).

Für die Untersuchung des Verbesserungspotenzials für die durch linguistische Extraktionsmuster erreichte Abdeckung soll hier ein anderer Weg begangen werden. Zunächst sollen zwar auch manuell zusätzliche Suchmuster spezifiziert werden, allerdings sollen diese aus der in Kapitel 3 referierten Literatur entnommen werden, anstatt auf Introspektion zu basieren. Bei einer solchen Spezifikation handelt es sich um ein schwieriges Unterfangen, da in der Literatur zumeist keine Valenzrahmen angegeben werden, diese im Sinne des hier verfolgten Ansatzes aber spezifiziert werden müssen, ohne dass Korpusbeispiele aus dem Entwicklungskorpus herangezogen werden können, da diese die betreffenden *knowledge patterns* gerade nicht in großem Umfang abdecken. Konkret werden die folgenden zusätzlichen Muster formuliert:

- *kommen_aus*: Term + kommt aus, Begriff/Name/Wort/Fachbegriff/Fachwort + Term + kommt aus,
- *begriff*: Begriff/Name/Wort/Fachbegriff/Fachwort + Term + steht (für)/bezeichnet/ bedeutet/meint/besagt/benennt,
- *umschreiben*: Begriff/Name + Term + umschreibt,
- *entstehen*: Term + entsteht,
- *umfassen*: Term + umfasst/beinhaltet/enthält,
- *fallen_unter*: Term + fällt unter, fällt unter + Term,
- *art_sein*: Art/Typ/Beispiel/Wort/Begriff/Fachbegriff + Präposition + Term + sein,
- *bekannt_sein*: Term + ist + bekannt,

- *gelten*: Term + gilt als, gilt als + Term,
- *definieren*: Term + wird + definiert, Term + definiert + man,
- *charakterisieren*: Term + ist + charakterisiert + durch,
- *auszeichnen*: Term + zeichnet + sich + aus,
- *zusammensetzen*: Term + setzt + sich + zusammen,
- *gebrauchen*: Term + wird + gebraucht, Term + gebraucht + man,
- *also*: Term + , + also/sprich/kurz/nämlich/sozusagen/z.B./d.h./u.a. + finites Verb im Tempus Präsens.

Mit diesen zusätzlichen Mustern wird folglich versucht, einen beträchtlichen Teil der in Tabelle 4 (Kapitel 3) referierten deutschen Extraktionsmuster zu implementieren. Allerdings erweist sich der Beitrag dieser zusätzlichen Muster – ähnlich wie bei Walter – als durchgängig gering. So werden für das Sub-EKD mit diesen Mustern nur sehr wenige zusätzliche Sätze extrahiert, von denen allerdings kein einziger ein bisher nicht extrahierter KRC ist. Für das Sub-TKD können immerhin acht neue KRCs extrahiert werden, allerdings nur um den Preis einer zusätzlichen Aufnahme falscher Positive. Tabelle 68 zeigt die Ergebnisse im Überblick.

Performanz	**Sub-EKD**	**Sub-TKD**
Extraktion mit zusätzlichen Mustern	7 zusätzliche falsche Positive, 5 bekannte KRCs, 3 schon vorher extrahierte falsche Positive	12 zusätzliche falsche Positive, 8 neue KRCs, 1 bekannter KRC, 1 schon vorher extrahierter falscher Positiv
neue Präzision	$\frac{231\ KRCs}{461\ alte\ Sätze + 7\ fP} = 0{,}4936$	$\frac{135\ KRCs + 8\ neue\ KRCs}{249\ alte\ Sätze + 12\ fP + 8\ neue\ KRCs} = 0{,}5316$
neue Abdeckung	0,6814 (231 KRCs)	$\frac{135\ KRCs + 8\ neue\ KRCs}{287} = 0{,}4983$
neues f-Maß	0,5725	0,5144

Tabelle 68: Performanz des deutschen Extraktionssystems mit zusätzlichen Suchmustern aus der Literatur

Der Tabelle ist zu entnehmen, dass zwar – angesichts der stark begrenzten Ressourcen, die hier wie im Falle Walters für die Spezifikation zusätzlicher Extraktionsmuster eingesetzt werden konnten, keineswegs überraschend – nur wenige zusätzliche KRC-Kandidaten extrahiert werden können. Im Fall des Sub-EKD ergibt sich dabei aufgrund der dabei zusätzlich aufgenommenen falschen Positive ein (sehr geringer) Abfall der Gesamtleistung, im Fall des Sub-TKD können dagegen Abdeckung und f-Maß geringfügig gesteigert werden. Werden im Ergebnis dieser Untersuchung nur diejenigen Muster in die Gesamtmenge aller Extraktionsmuster übernommen, die zumindest aus dem Sub-TKD valide KRCs zu extrahieren vermögen, ergibt sich für das Sub-EKD ein f-Maß von 0,57 (Präzision von 50% und Abdeckung von 68%) und für das Sub-TKD ein Maß von 0,52 (Präzision von 54% und Abdeckung von 50%).

Zu prüfen ist nun noch das Verbesserungspotential durch Relaxierung bzw. Modifikation bereits existierender Muster. Hierfür können beispielsweise die Abdeckungsfehler auf dem Sub-TKD herangezogen werden, d. h. es kann überprüft werden, aus welchen Gründen bestimmte KRCs nicht aus diesem Korpus extrahiert werden. Zu beachten ist dabei allerdings, dass eine solche Prüfung nicht in jedem Fall zweifelsfreie Ergebnisse erbringen kann, da u. U. mehrere, einander bedingende Ursachen für die Nicht-Extraktion eines validen Kontexts vorliegen können bzw. angesichts der

Komplexität der Korpusannotation und der darauf aufbauenden Extraktionsmuster die eigentliche Ursache für ein *miss* nicht immer offensichtlich ist. Aus diesem Grund werden an dieser Stelle nur Tendenzen, nicht aber konkrete Zahlen angegeben. Die Analyse zeigt:

- Die häufigste, etwa für ein Drittel der nicht extrahierten KRCs verantwortliche Fehlerursache sind fehlende Extraktionsmuster. Allerdings – dies zeigen die Beispiele 104) bis 107) – handelt es sich bei den Prädikaten bzw. den das Definiendum einführenden Strukturen in den nicht extrahierten KRCs des Sub-TKD um recht spezifische Verben bzw. nur schwer von anderen Strukturen abgrenzbare Phänomene, die weder durch allgemeine Musterkataloge noch durch die beispielsweise von Walter angewandte Methode der Introspektion identifiziert werden können. Vielmehr scheinen in diesen Beispielen in Bestätigung der im vorigen Kapitel untersuchten quantitativen Eigenschaften wissenshaltiger Kontexte andere sprachliche Mittel (z. B. Parenthesen, Pluralverwendung) zum Einsatz zu kommen, um zu verdeutlichen, dass begriffliche Informationen vermittelt werden[130].

104) Akkumulatoren (Batterien) speichern Strom elektrochemisch .

105) Aufwindkraftwerke erzeugen durch eine geeignete Konstruktion (ein umgekehrter Trichter) einen starken thermischen Aufwind , der sich mit Turbinen nutzen lässt .

106) Bei großen Kraftwerksgeneratoren sitzt noch ein weiterer Gleichstromgenerator auf der gemeinsamen Welle, die Hilfserregermaschine.

107) Das heißt , es fehlt an Erfindungshöhe , wenn man von einem Fachmann erwarten kann , dass er , ausgehend vom Stand der Technik , auf diese Lösung alsbald und mit einem zumutbaren Aufwand gekommen wäre , ohne erfinderisch tätig zu werden .

Darüber hinaus zeigt die Analyse der falschen Negative im Sub-TKD aber auch:

- Eine etwa ebenso häufige Ursache für ein *miss* sind Fehler bei der linguistischen Annotation des Korpus‘ sowie bei der Termerkennung, die ja als Grundlage für die musterbasierte Extraktion dient. Erstere Feststellung deckt sich mit der Untersuchung Walters (ebd.: 194 ff.), der eine sehr ausführliche Fehleranalyse leistet und insbesondere Abdeckungsfehler (also falsche Negative bzw. *misses*) auf diesen Fehlertyp zurückführt. Annotationsfehler können dabei auf unterschiedlichen Vorverarbeitungsstufen (z. B. Satzgrenzenerkennung, syntaktische Annotation etc.) auftreten bzw. aufeinander aufbauen. Termerkennungsfehler beruhen dagegen auf der inkonsistenten Lemmatisierung der Definienda bzw. Zieltermini durch das automatische Analysewerkzeug. Geschuldet sind diese Fehler nicht nur der zwar sehr guten, aber eben nicht perfekten Performanz der eingesetzten Werkzeuge, sondern auch der sprachlichen und inhaltlichen Komplexität der untersuchten Datensätze.
- Eine etwas geringere Häufigkeit weisen mit einem Anteil von ca. 20% *misses* aufgrund von zu eng formulierten existierenden Suchmustern auf, wobei es sich in etwa der Hälfte der Fälle um solche Muster handelt, die bereits unpräzise sind.
- Eine geringe Anzahl falscher Negative geht schließlich auf die zur Präzisionsverbesserung eingesetzten Filter zurück.

[130] Die Beispiele werden so dargestellt, wie sie auch in den nicht-annotierten Korpora auftreten, d. h. insbesondere mit den dort auftretenden überzähligen Leerzeichen.

Insgesamt kann das Verbesserungspotential durch die Relaxierung von Mustern mithin als eher gering eingeschätzt werden, wenn davon abgesehen wird, ohnehin schon unpräzise Muster zu erweitern. Die von Walter angewandte Methode des *bootstrapping* erscheint allerdings auch nicht als vielversprechend, da die Goldstandard-Korpora – insbesondere das multidisziplinäre Sub-TKD – als für einen solchen Ansatz zu klein anzusehen sind. Des Weiteren – dies zeigen auch die obigen Beispiele – kann nicht für jeden KRC ein Lemmapaar Definiendum-*genus proximum* identifiziert werden, so dass die Nützlichkeit der Methode von vornherein eingeschränkt scheint. Um jedoch immerhin eine Abschätzung des Verbesserungspotenzials durch die Erweiterung bzw. Modifikation von Suchmustern geben zu können, sollen aufbauend auf der Analyse der *misses* für das Sub-TKD einige Mustermodifikationen vorgenommen werden. Dabei werden der bisherigen Strategie folgend nur solche Muster verändert, die für beide Korpora zu Performanzverbesserungen bzw. höchstens geringfügigen –verschlechterungen für eines der Korpora führen. Konkret werden folgende Modifikationen implementiert:

- Für das *knowledge pattern bezeichnen* wird eine Variante implementiert, in der das Definiendum Argument der Präposition „als“ ist. Für das *knowledge pattern zählen_zu* wird dagegen die inverse Modifikation umgesetzt.
- Für das *knowledge pattern nennen* wird eine Variante mit dem Hilfsverb „sein“ hinzugefügt.
- Für das *knowledge pattern verwenden* werden Hilfsverbmodifikationen vorgenommen.
- Für die *knowledge patterns beschreiben* und *herstellen* werden Varianten mit einem vor das Definiendum vorgeschalteten Nomen implementiert.
- Für das *knowledge pattern benutzen* wird eine Variante mit dem Hilfsverb „werden“ hinzugefügt.

Für das Sub-EKD können mit diesen Veränderungen keine positiven Effekte erzielt werden, die Präzision geht wiederum leicht zurück. Für das Sub-TKD kann die Abdeckung dagegen leicht verbessert werden. Tabelle 69 fasst die neuen Werte zusammen.

Evaluationsmaß	**Sub-EKD (alt)**	**Sub-EKD (neu)**	**Sub-TKD (alt)**	**Sub-TKD (neu)**
Präzision	0,4968 (465 Sätze)	0,4936 (468 Sätze)	0,5356 (267 Sätze)	0,5406 (283 Sätze)
Abdeckung	0,6814 (231 KRCs)	0,6814 (231 KRCs)	0,4983 (143 KRCs)	0,5331 (153 KRCs)
f-Maß	0,5746	0,5725	0,5162	0,5368

Tabelle 69: Ergebnisse der Maßnahmen zur Verbesserung der Abdeckung für das deutsche Extraktionssystem

6.2.3.2 Russisch

6.2.3.2.1 Baseline

Die mit Hilfe des einfachen russischen *baseline*-Extraktionssystems erzielten Ergebnisse können nun ebenfalls ausgewertet werden. Wie im Falle des Deutschen werden dabei solche Treffer, die Goldstandard-KRCs stark ähneln, dennoch nicht als *true positives* gewertet, was insbesondere im Falle des Sub-EKR eine höhere Präzision erbringen würde. Als Bewertungsgrundlage werden 293 (Sub-EKR) bzw. 267 (Sub-TKR) extrahierbare KRCs herangezogen, die in den annotierten Goldstandard-Korpora anhand ihrer Satzidentifikationsnummern erkannt werden. Die Gründe für die Abweichungen zu den in Kapitel 4 genannten Zahlen sind dabei die gleichen wie im Fall des Deutschen. Auch hier werden zudem von unterschiedlichen Mustern mehrfach extrahierte Sätze für

die Berechnung der globalen Maße Präzision, Abdeckung und f-Maß nur einfach evaluiert. Zur Bewertung der Performanz der einzelnen Extraktionsmuster wird dagegen die Gesamtmenge der von einem Muster extrahierten Sätze herangezogen. Tabelle 70 fasst die Ergebnisse für das Russische zusammen, Abbildung 30 zeigt den entsprechenden Präzisions-Abdeckungs-Graphen.

Evaluationsmaß	**Sub-EKR**	**Sub-TKR**
Präzision	0,5678 (317 extrahierte Sätze)	0,3491 (169 extrahierte Sätze)
Abdeckung	0,6143 (180 extrahierte KRCs)	0,2210 (59 extrahierte KRCs)
f-Maß	0,5902	0,2706

Tabelle 70: Performanz des Baseline-Extraktionssystems auf den russischen Goldstandard-Korpora

Diese Ergebnisse weichen deutlich von den für das Deutsche erzielten Resultaten ab. Zwar wird auf dem Entwicklungskorpus eine im Vergleich zum Deutschen bessere Gesamtperformanz erreicht, jedoch kann der Leistungsabfall auf den Testdaten sowohl in Bezug auf die Präzision als auch auf die Abdeckung und dementsprechend auf die Gesamtperformanz nicht anders als dramatisch genannt werden. Eine Fehleranalyse zeigt dabei, dass für falsche Negative auf dem Testkorpus insbesondere – zu nahezu 50% – fehlende Extraktionsmuster verantwortlich sind. Bis zu 25% der Abdeckungsfehler gehen zudem jeweils auf nicht erkannte Termini oder zu restriktiv formulierte Suchmuster zurück. Nicht zuletzt sind auch Annotationsfehler bzw. inkonsistente *parser*-Ausgaben eine weitere Ursache für KRC-*misses*. Damit zeigt sich, dass im Vergleich zum Deutschen die Performanz der in beiden Korpora (Sub-EKR und Sub-TKR) gleichermaßen genutzten *knowledge patterns* deutlich geringer ist – die russischen Muster des Entwicklungskorpus‘ scheinen nur mit geringem Erfolg auf das Testkorpus übertragbar.

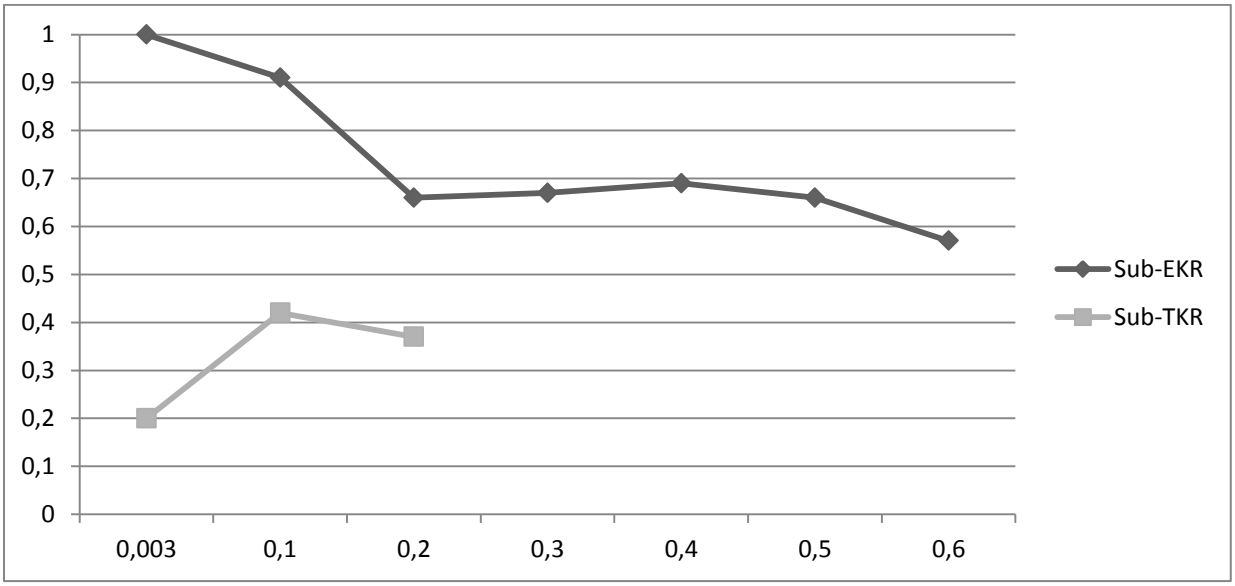


Abbildung 30: Präzisions-Abdeckungs-Graph für das russische Baseline-Extraktionssystem

Eine Ursache hierfür scheint zu sein, dass insbesondere generische Kopula (wie im Deutschen das Verb „sein“), die in der Regel viele KRCs extrahieren, im russischen Entwicklungskorpus selten auftreten und entsprechend nicht berücksichtigt wurden, wohingegen im Testkorpus durchaus eine Reihe entsprechender Formulierungen erscheint. Darüber hinaus verwendet das Russische zusätzlich zu variablen Prädikaten, mit deren Hilfe Definiendum und Definiens verbunden werden, in der Tat eine Vielzahl linguistischer (grammatischer und syntaktischer) Strukturen, die durch die in der

Literatur für andere Sprachen beschriebenen und in der Regel recht einfachen Strukturen ebenso wenig abgedeckt werden wie durch die kleine Studie auf Grundlage des Sub-EKR, vgl. die aus dem Sub-TKR nicht extrahierten folgenden Beispiele:

108) Основными видами термической обработки являются отжиг, нормализация, закалка и отпуск.
[*Übersetzung*: Die Grundformen der Wärmebehandlung sind Glühen, Normalglühen, Abschrecken und Anlassen.]

109) Точечная сварка в зависимости от расположения электродов по отношению к свариваемым заготовкам может быть двусторонней и односторонней.
[*Übersetzung*: Das Punktschweißen kann in Abhängigkeit von der Position der Elektroden zu dem zu schweißenden Werkstück sowohl ein- als auch zweiseitig sein.]

110) В процессе термической обработки в металле происходят превращения, в результате которых улучшаются прочностные свойства, устраняются напряжения, повышается износостойкость, придаются металлу требуемая твердость, хорошая обрабатываемость режущим инструментом и др.
[*Übersetzung*: Während der Wärmebehandlung finden im Metall Veränderungen statt, in deren Ergebnis die Festigkeitseigenschaften verbessert und Spannungen beseitigt werden können, die Verschleißfestigkeit wird erhöht und das Metall erhält die notwendige Härte, lässt sich gut schneiden etc.]

111) Закалкой придают деталям высокую твердость, прочность, износостойкость.
[*Übersetzung:* Durch das Abschrecken erhalten die Werkstücke die erforderliche Härte, Haltbarkeit und Verschleißfestigkeit.]

Allein diese wenigen Beispiele zeigen, dass das Russische für die Realisierung einfacher Kopula neben dem als Muster implementierten Verb *представлять собой* und den verblosen Bindestrichkonstruktionen eine Reihe weiterer (meist imperfektiver) Verben einsetzt, die ggf. auch mit einem Modalverb kombiniert sein können. Waren für das Deutsche Muster mit der Präposition „bei“ als Einleitung von das Definiendum thematisierenden Situativkomplementen entwickelt worden, könnte eine ähnliche Funktionen möglicherweise der russischen Präpositionen *при* zugeschrieben werden, allerdings verfügt das Russische darüber hinaus über ein produktiv erweiterbares Spektrum weiterer Präpositional- bzw. Konnektorphrasen, z. B. *в процессе* (wörtlich: im Prozess + Genitiv, „während“), *в ходе* (wörtlich: im Verlauf + Genitiv, „während“), *в результате* (wörtlich: im Ergebnis + Genitiv, „durch“) etc., die zu ähnlichen Zwecken eingesetzt werden können. Ähnliche Phänomene lassen sich auch für Modalangaben beobachten (z. B. *для -> в целях*, „für“). Lexikalische Vielfalt zeigt sich darüber hinaus in der Verwendung unterschiedlichster Nomina als *genus proximum* (z. B. *тип*, „Typ“, *вид*, „Art“, „Typ“, *разновидность*, „Art“, „Typ“, „Unterart“) sowie in der Erweiterung der Liste der verwendeten *knowledge patterns* durch Präfigierung von Verben bei gleichzeitiger Beibehaltung der Aspekteigenschaften (z. B. *разделяться -> подразделяться,* „unterteilt werden“). Darüber hinaus finden sich unter den nicht extrahierten KRCs des Sub-TKR zahlreiche verblose Bindestrich-Konstruktionen, die durch die – relativ eng – formulierten *tire*-Muster nicht abgedeckt werden, deren Extrahierbarkeit allerdings – zumindest dann, wenn eine mindestens mittlere Präzision angestrebt werden soll – grundsätzlich zweifelhaft

scheint, da Bindestriche in Korpora in einer Vielzahl unterschiedlichster Formate und Funktionen auftreten können. Nicht zuletzt scheinen bei der Musterformulierung auch Varianten von Mustern z. B. mit abhängigem Definiendum nicht in ausreichendem Maße berücksichtigt worden zu sein.

Erschwert wird die Extraktion wissenshaltiger Kontexte zudem durch den Umstand inkonsistenter bzw. fehlerhafter Annotationen durch die verwendeten Annotationswerkzeuge und eine stärkere Variation grammatischer Strukturen, als dies zumindest anhand des kleinen Sub-TKR absehbar war. Beispielsweise waren für ähnliche Strukturen wie diejenige in Beispiel 111) die Extraktionsmuster *agent1* und *agent2* entwickelt worden. Diese Muster suchen, wie in Anhang 11 dargestellt, nach Termini im Instrumental („@morph[contains(., ‚i')]“) und mit der syntaktischen Funktion *агент* („Agens“, vgl. Tabelle 36, Abschnitt 4.4.2.2.2), denen ein Verb unmittelbar vorausgeht. Dieses Muster extrahiert beispielsweise den in 112) dargestellten KRC:

112) Контроль за нераспространением ядерного оружия осуществляется с помощью Международного агентства по атомной энергии (МАГАТЭ, IAEA — International Atomic Energy Agency), с которым каждый участник Договора, не обладающий ядерным оружием, обязан заключить соответствующее соглашение.
[*Übersetzung:* Die Kontrolle über die Nichtverbreitung von Atomwaffen wird durch die Internationale Atomenergie-Organisation (IAEO, engl. IAEA — International Atomic Energy Agency) realisiert, mit der jeder Vertragspartner[131], der nicht über Atomwaffen verfügt, ein entsprechendes Abkommen schließen muss.

Das Beispiel 111) wiederum wird durch die fraglichen Muster nicht extrahiert, da hier das Agens im Instrumental (*закалкой*, „durch das Abschrecken“) nicht nur vor dem Verb positioniert ist, sondern auch durch den *parser* nicht als Agens, sondern als „ROOT“ annotiert wurde. Die bereits angesprochene Möglichkeit der syntaktisch-positionellen Variation von Satzbestandteilen trifft in diesem Beispiel also mit fehlerhaften bzw. zumindest inkonsistenten Annotationen zusammen und erschwert damit die KRC-Extraktion. Zusätzlich zu diesen Aspekten gelten auch die bereits für das Deutsche gemachten Bemerkungen zur Übertragbarkeit von Extraktionsmustern vom Entwicklungs- auf das Testkorpus, d. h. die Performanz individueller Extraktionsmuster schwankt beim Wechsel auf das Testkorpus relativ stark und eine recht große Anzahl von Extraktionsmustern trägt gar nicht zur Extraktion von KRCs aus dem Testkorpus bei, wodurch nicht zuletzt auch die Präzision der Extraktion leidet. Da im Hinblick auf das Testkorpus insbesondere das Erreichen einer akzeptablen Abdeckung als vorrangiges Problem erscheint, sollen in den folgenden Abschnitten Ansätze im Wesentlichen zur Verbesserung dieser Größe überprüft und bewertet werden.

6.2.3.2.2 Term-Erkennung

Die Analyse der Abdeckungsfehler im vorigen Abschnitt hat ergeben, dass im Vergleich zum Deutschen die Term-Erkennung aufgrund inkonsistenter Lemmatisierung häufig fehlschlägt und KRC-Kandidaten entsprechend nicht extrahiert werden können. Waren für das Deutsche Term-Erkennungs- und andere Annotationsfehler gemeinsam und etwa zu gleichen Teilen für ca. ein Drittel der falschen Negative verantwortlich, können für das Russische bis zu 25% der *misses* auf nicht erkannte Termini bzw. Definienda zurückgeführt werden.

[131] Gemeint ist hiermit der Atomwaffensperrvertrag.

Die Term-Erkennung war in Kapitel 4 kurz beschrieben worden: Lemmatisierte Zieltermini waren dort auf ebenfalls lemmatisierte Einheiten – die XML-Lemma-Attribute jedes *tokens* – abgebildet worden. Die in Kapitel 4 durchgeführte knappe Evaluation dieses Annotationsschrittes deutete dabei bereits darauf hin, dass insbesondere für das Sub-TKR dieses Verfahren der Termerkennung nur eine teilweise Identifikation der Termini im Korpus leistet (vgl. Tabelle 38, Abschnitt 4.4.3).
Um zu bewerten, inwieweit eine bessere Term-Erkennung zu einer Verbesserung der Abdeckung durch das musterbasierte Extraktionsverfahren beitragen kann, soll nun versucht werden, die Term-Erkennung auf dem Sub-TKR zu verbessern. Dabei wird wie folgt vorgegangen:

- Russische Mehrworttermini unterliegen zwar produktiven, aber gleichzeitig ziemlich regelhaften Wortbildungsmustern. Darüber hinaus sind die Flexionsparadigmen russischer Nominalphrasen zwar in wesentlichen Teilen ambig – daher das Problem inkonsistenter und falscher automatischer Lemmatisierungen –, aber auch weitgehend regelhaft.
- Eine Vollformenliste russischer Ein- und Mehrworttermini lässt sich daher mit vertretbarem Aufwand semi-automatisch generieren.
- Eine solche Liste soll für die Zieltermini des Sub-TKR mittels eines einfachen Perl-Skripts generiert und manuell korrigiert werden, wobei auch Varianten für ambige Schreibungen (z. B. *e*/*ë*) berücksichtigt werden.
- Diese Liste soll alsdann auf das in XML kodierte Sub-TKR angewendet werden, wobei die Terminus-Vollformen nicht wie bisher mit den Lemma-Attributen der einzelnen Wort-*tokens* abgeglichen werden, sondern – unter Ignorierung von Groß- und Kleinschreibung – mit den Wort-*tokens* selbst. Hierbei werden auch Normalisierungsaspekte (z. B. das Entfernen von Markierungen des Wortakzents im Sinne einer weiteren Vereinheitlichung der Schreibungen als Grundlage für ein erfolgreiches *matching*) berücksichtigt.

Die auf diese Weise neu entstehende Fassung des Korpus‘ enthält insgesamt 4295 *tokens* mit dem Attribut „@isterm=‘true‘“. Die ältere Fassung des Korpus‘ umfasste lediglich 3949 *tokens* mit diesem Merkmal. Das angepasste Term-Erkennungs-Verfahren scheint also erfolgreich gewesen zu sein. Tabelle 71 stellt nun die im Ergebnis dieser Modifikation auf dem Sub-TKR erreichte Performanz den älteren Werten gegenüber.

Evaluationsmaß	**Sub-TKR (alt)**	**Sub-TKR (neu)**
Präzision	0,3491 (169 extrahierte Sätze)	0,4000 (190 extrahierte Sätze)
Abdeckung	0,2210 (59 extrahierte KRCs)	0,2846 (76 extrahierte KRCs)
f-Maß	0,2706	0,3326

Tabelle 71: Performanz des russischen Extraktionssystems auf dem Sub-TKR nach verbesserter Term-Erkennung

Die Tabelle zeigt, dass tatsächlich einige zusätzliche KRCs extrahiert werden können, wobei falsche Positive nur in geringem Maße generiert werden, was sich insgesamt in einer Verbesserung der Gesamtperformanz niederschlägt. Der Anteil der nicht extrahierten KRCs ist jedoch nach wie vor sehr groß und die Leistungssteigerung aufgrund der verbesserten Term-Erkennung liegt deutlich unter der als Maximum angenommenen Größe von 25%, so dass in den verbleibenden Fällen immer noch andere Fehlerursachen vorliegen müssen, beispielsweise Annotationsfehler oder fehlende bzw. zu restriktiv formulierte Extraktionsmuster. Tatsächlich gehen etwa 70% der für diese Gruppe verbleibenden *misses* trotz verbesserter Term-Erkennung auf fehlende oder zu restriktiv formulierte Extraktionsmuster zurück, inkonsistente oder fehlerhafte *parser*-Ausgaben sowie ungewöhnliche Realisierungsvarianten der Zieltermini in den KRCs haben dagegen nur einen geringen Anteil an den verbleibenden Abdeckungsfehlern.

6.2.3.2.3 Zusätzliche Extraktionsmuster

Eine weitere Möglichkeit für die Verbesserung der Abdeckung ist wie bereits im Falle des Deutschen die Formulierung zusätzlicher Extraktionsmuster. Auch hier soll dabei theoriebasiert vorgegangen werden, womit gemeint ist, dass in der Literatur für andere Sprachen beschriebene Muster (vgl. Tabelle 4, Abschnitt 3.2.3) ins Russische „übersetzt“ werden sollen. Für die Extraktion verwendet werden dabei wiederum nur solche Muster, die für das Sub-TKR eine Performanzsteigerung bewirken, ohne die Leistung des Sub-EKR allzu stark herabzusetzen. Konkret werden die folgenden zusätzlichen Muster implementiert:

- *образоваться*: Term + образоваться/образуется/образуются,
- *относиться*: Term + относиться/относится/относятся + к,
- *тип*: тип/разновидность/вид + Term,
- *понимать*: под + Term + понимать/понимает/понимают/пониматься /понимается/понимаются/подразумевать/подразумевают/подразумеваться/ подразумеваются,
- *являться*: Term + являться/является/являются + Nomen im Instrumental,
- *представлять*: Term + представлять/представляет/представляют,
- *известен*: Term + token[contains(., 'извест')].

Mit diesen wenigen zusätzlichen Mustern kann für das Sub-TKR tatsächlich eine deutliche Leistungssteigerung erzielt werden. Tabelle 72 fasst die neuen Performanzwerte zusammen. Es zeigt sich, dass die Leistung des Extraktionssystems auf dem Entwicklungskorpus zwar leicht zurückgeht, die Performanz auf dem Testkorpus‘ dank eines sehr deutlichen Anstiegs der Abdeckung – bei allerdings leicht verminderter Präzision – jedoch deutlich angehoben werden kann.

Evaluationsmaß	**Sub-EKR (alt)**	**Sub-EKR (neu)**	**Sub-TKR (alt)**	**Sub-TKR (neu)**
Präzision	0,5678 (317 extrahierte Sätze)	0,5403 (335 extrahierte Sätze)	0,4000 (190 extrahierte Sätze)	0,3897 (272 extrahierte Sätze)
Abdeckung	0,6143 (180 extrahierte KRCs)	0,6177 (181 extrahierte KRCs)	0,2846 (76 extrahierte KRCs)	0,3970 (106 extrahierte KRCs)
f-Maß	0,5902	0,5764	0,3326	0,3933

Tabelle 72: Performanz des russischen Extraktionssystems mit zusätzlichen theoriebasierten Extraktionsmustern

6.2.3.2.4 Punktuelle Präzisionsverbesserung

Als vorerst letzter möglicher Optimierungsschritt können Möglichkeiten zur Verbesserung der erreichten Präzision diskutiert werden. Wie im Falle des Deutschen bieten sich auch hier zwei Ansätze an, und zwar einerseits die Formulierung von Ausnahmeregeln, bei deren Vorliegen ein KRC-Kandidat nicht extrahiert werden soll, sowie andererseits die Anpassung bestimmter Extraktionsmuster (auf das Löschen unpräziser Muster soll im Hinblick auf die immer noch niedrige Abdeckung auf dem Sub-TKR verzichtet werden). Beide Ansätze sollen hier wiederum so umgesetzt werden, dass nur diejenigen Modifikationen implementiert werden, die auf dem Entwicklungskorpus keinen deutlichen Leistungsabfall und auf dem Testkorpus eine Leistungssteigerung bewirken. Im Ergebnis werden zusätzlich zu den bereits verwendeten (vgl. Abschnitt 6.2.2.2) folgende Ausnahmeregeln implementiert:

- keine Extraktion von Sätzen, die mit einem Demonstrativpronomen wie beispielsweise *этот* beginnen,

- keine Extraktion von Sätzen, die Partikeln enthalten,
- keine Extraktion von Sätzen, die mit einer Konjunktion beginnen,
- keine Extraktion von Sätzen, die mit einer der Phrasen *Таким образом* („Auf diese Weise“), *Кроме того* („Außerdem“), *При этом* („Dabei“) oder *Для этого* („Dafür“) beginnen oder welche die Phrase *при этом* enthalten.

Darüber hinaus werden zu den Extraktionsmustern folgende Einschränkungen hinzugefügt:

- Das Muster *agent2* darf nur dann KRC-Kandidaten extrahieren, wenn sich das fragliche Nomen im Instrumental auf ein vorangehendes Verb bezieht, das *keine* Partizipialform („[substring(@morph, 3, 1)!='p']“), also möglichst ein konjugiertes Vollverb ist.
- Das Muster *parenthese1* darf nur dann KRC-Kandidaten extrahieren, wenn diese keine Verbformen im Präteritum („token[@pos='V'][substring(@morph, 4, 1)='s']“) enthalten.
- Das im vorigen Abschnitt neu hinzugefügte Muster *относиться* („gehören zu“) wurde um eine Regel ergänzt, die ausschließt, dass als präpositionales Objekt Pronomen auftreten können.
- Das ebenfalls im letzten Abschnitt neu hinzugefügte Muster *тип* („Typ“) wurde ebenfalls um bestimmte lexikalische Ausnahmeregeln ergänzt, um Homonymie-Effekte zu reduzieren.

Mittels dieser bestenfalls als punktuell zu charakterisierenden Modifikationen können die in Tabelle 73 zusammengefassten Performanzverbesserungen erreicht werden. Abbildung 31 schließlich stellt die sich nun ergebenden Präzisions-Abdeckungs-Graphen den entsprechenden Graphen auf Grundlage des *baseline*-Systems (Abbildung 30, Abschnitt 6.2.3.2.1) gegenüber.

Evaluationsmaß	Sub-EKR (alt)	Sub-EKR (neu)	Sub-TKR (alt)	Sub-TKR (neu)
Präzision	0,5403 (335 extrahierte Sätze)	0,5897 (290 extrahierte Sätze)	0,3897 (272 extrahierte Sätze)	0,4518 (228 extrahierte Sätze)
Abdeckung	0,6177 (181 extrahierte KRCs)	0,5836 (171 extrahierte KRCs)	0,3970 (106 extrahierte KRCs)	0,3858 (103 extrahierte KRCs)
f-Maß	0,5764	0,5866	0,3933	0,4162

Tabelle 73: Performanz des russischen Extraktionssystems nach Präzisionssteigerung

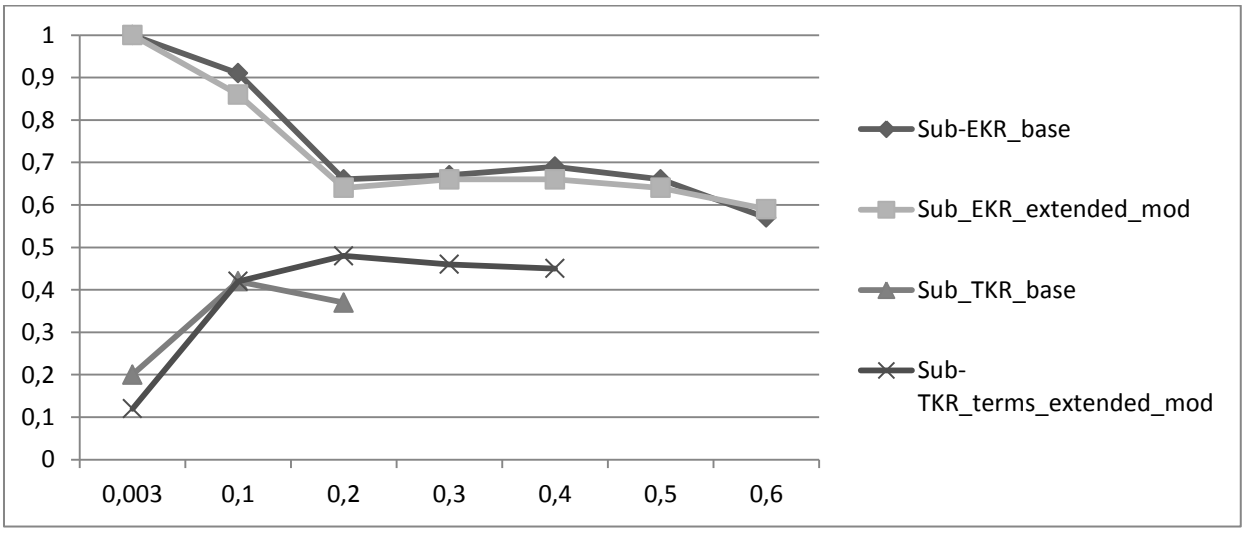


Abbildung 31: Vergleich der Präzisions-Abdeckungs-Graphen des russischen Baseline-Systems (_base) mit den entsprechenden Graphen nach der Präzisions- und Abdeckungsoptimierung (_extended_mod)

Abbildung und Tabelle kann entnommen werden, dass die beschriebenen Anpassungen in Bezug auf das Sub-EKR lediglich einen zu vernachlässigenden Einfluss haben. In Bezug auf das Sub-TKR können

dagegen sowohl Präzision als auch Abdeckung deutlich verbessert werden, wobei allerdings der Leistungsabfall beim Übergang vom Entwicklungs- zum Testkorpus nach wie vor beträchtlich ist. Tabelle 74 fasst wiederum analog zur Darstellung für das Deutsche die Präzision der individuellen Suchmuster auf beiden Korpora zusammen[132].

Muster	**Präzision Sub-EKR**	**Präzision Sub-TKR**
agent1	*0,57*	*0,00*
agent2	0,31	0,35
parenthese1	0,33	0,44
prep1	*0,38*	*1,00*
tire1	**1,00**	**1,00**
tire2	*1,00*	-
tire3	*1,00*	-
tire4	**1,00**	**1,00**
включить1	*1,00*	-
включить2	*0,56*	-
воздействовать	*0,75*	-
делиться	**1,00**	**0,50**
заключаться	*1,00*	-
известен	*0,00*	*1,00*
иметь1	0,21	0,29
использовать1	*0,50*	-
использовать2	0,43	0,43
использоваться	*1,00*	*0,13*
корректировать	*0,50*	-
который	0,38	0,25
называть	*0,00*	*0,75*
называться	**0,50**	**0,67**
обеспечивать1	0,23	0,33
обеспечивать3	0,47	-
оборудоваться	*1,00*	-
образоваться	-	0,20
осуществляться1	*0,40*	*0,50*
осуществляться2	*1,00*	-
относиться	*0,00*	*0,50*
передавать	*1,00*	-
поддерживать	*1,00*	*0,00*
позволять	**0,63**	**0,67**
понимать	-	*1,00*
предназначен	**0,93**	**0,71**
представлять	**0,83**	**1,00**
представлять1	**0,83**	**1,00**
представлять2	**1,00**	**1,00**
приводиться	*0,50*	*0,00*
применять1	0,00	0,38
применять2	0,25	0,00
применяться1	*0,60*	*0,38*

[132] Die Muster *оборудован* und *применятся2* extrahieren aufgrund der implementierten Filter keine KRC-Kandidaten mehr und entfallen daher in der weiteren Darstellung.

различать	**1,00**	**0,71**
расположен	*0,36*	*0,50*
служить1	**0,73**	**0,50**
служить3	**0,57**	**1,00**
служить5	*1,00*	-
снижать1	*1,00*	-
снижать2	*1,00*	-
состоять1	**0,85**	**0,83**
состоять3	**0,88**	**1,00**
состоять4	*1,00*	-
состоять5	**1,00**	**1,00**
тип	0,29	0,48
управлять	*1,00*	-
устанавлиаться	*0,67*	*0,00*
устанавливать1	*0,60*	-
установить1	0,33	-
установить2	**0,50**	**0,50**
являться	0,00	0,26
durchschnittliche Präzision	**0,68**	**0,56**

Tabelle 74: Präzision der individuellen Suchmuster auf den russischen Korpora

Auch hier zeigt sich, dass die Leistungsfähigkeit für das Gros der einzelnen Muster beim Übergang vom Sub-EKR auf das Sub-TKR stark schwankt: Während einige wenige Extraktionsmuster auf beiden Korpora annähernd stabile Leistungen erbringen, erweisen sich andere Muster als korpusspezifisch – sie tragen zur Extraktion von KRCs aus dem Sub-TKR nicht bei, weisen in der Regel aber auf dem Entwicklungskorpus hohe Präzisionswerte auf. Auch hier kann mithin von einer Überanpassung der Muster an das Entwicklungskorpus ausgegangen werden. Allerdings ist der Abfall in der durchschnittlichen Musterpräzision geringer als im Deutschen (Tabelle 66, Abschnitt 6.2.3.1.1), was möglicherweise auf die bereits durchgeführte Optimierung zurückzuführen ist[133]. Für diese Erklärung spricht auch die in Abbildung 31 verdeutlichte Präzisionssteigerung für das Sub-TKR im Ergebnis der beschriebenen Optimierungsschritte. Wiederum andere Muster zeigen deutlich unterschiedliche Leistungen, während nur wenige Extraktionsmuster durchgängig eine geringe Präzision aufweisen. Bei den auf beiden Korpora durchgängig guten Mustern handelt es sich wie bereits im Falle des Deutschen um eher generische Muster, die Mustern aus der Literatur ähneln, allein jedoch schwerlich eine angemessene Abdeckung erzielen können.

6.2.4 Diskussion

In den vorstehenden Abschnitten wurden die Implementierung, Optimierung und Evaluation musterbasierter Systeme für die Extraktion wissenshaltiger Kontexte aus den deutschen und russischen Goldstandard-Korpora beschrieben. Dabei stand allerdings weniger die Entwicklung konkreter Optimierungsalgorithmen im Vordergrund – weshalb nur punktuelle und mit hoher Wahrscheinlichkeit nicht oder nur in Teilen auf andere Datensätze übertragbare Optimierungsschritte beschrieben wurden – als vielmehr eine allgemeine Abschätzung und

[133] Tatsächlich beträgt die durchschnittliche Musterpräzision für das *baseline*-System auf dem Sub-TKR lediglich 0,45.

Bewertung der Leistungsfähigkeit musterbasierter Ansätze für die terminographisch ausgerichtete KRC-Extraktion. Konkret waren am Beginn dieses Kapitels die folgenden Fragen aufgeworfen worden:

- Ergeben sich bei der Extraktion russischer wissenshaltiger Kontexte besondere Schwierigkeiten, beispielsweise aufgrund der reichen Flexionsparadigmen des Russischen und seiner im Vergleich zum Deutschen noch freieren Wortfolge?
- Ist ein syntaxbasierter Ansatz der Musterformulierung auch für das Russische geeignet?
- Welche Performanz kann mit einem musterbasierten Extraktionsansatz generell erreicht werden?

Bezugnehmend auf die letzte dieser drei Fragen kann zunächst festgestellt werden, dass musterbasierte Ansätze zumindest auf den hier untersuchten Daten, wie bereits im Zusammenhang mit der Durchsicht der relevanten Fachliteratur vermutet, eine mittlere Performanz erbringen: Die im Ergebnis aller Optimierungsschritte für jedes Korpus erreichten f-Maße liegen zwischen 0,4 und 0,6 und damit – auf die Unmöglichkeit direkter Vergleiche sei allerdings nochmals hingewiesen – immerhin im oberen Bereich der in der Literatur für ähnliche Experimente genannten Performanzwerte, wobei das russische Testkorpus deutlich hinter den anderen Korpora zurückbleibt. Die Behauptung einer grundsätzlichen Überlegenheit von ML-Ansätzen gegenüber musterbasierten Verfahren für die KRC-Extraktion kann mithin auf Grundlage der hier beschriebenen Experimente und deren Evaluation nicht unterstützt werden. Deutlich wird auch, dass auf dem russischen Entwicklungskorpus konstant das höchste und im Ergebnis der beschriebenen Modifikationen kaum veränderte f-Maß erreicht wurde, wohingegen die Performanz der musterbasierten Extraktion auf den verbleibenden zwei Korpora leicht gesteigert werden konnte. Abbildung 32 stellt diese Zusammenhänge noch einmal im Überblick dar.

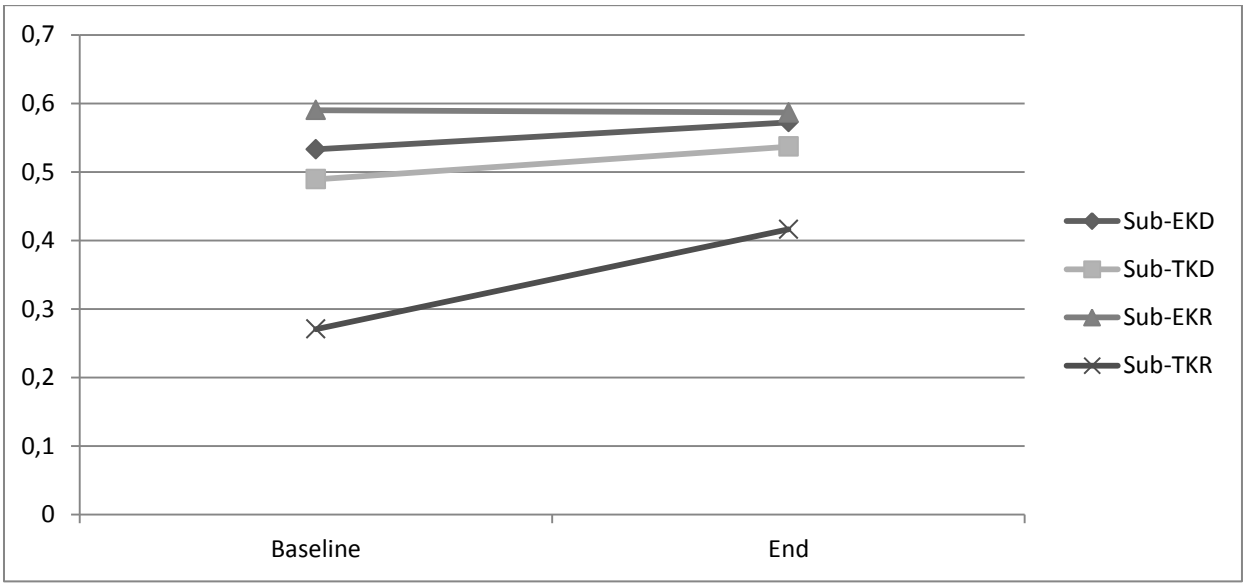


Abbildung 32: f-Maße für die musterbasierten Extraktionssysteme auf den einzelnen Korpora vor (Baseline) und nach der Optimierung (End)

Die Experimente machen zudem deutlich, dass Mustermengen sich sowohl in Richtung einer maximalen Präzision als auch auf eine maximale Abdeckung hin optimieren lassen, wobei jedoch in der Regel die Gesamtperformanz, d. h. das f-Maß, leidet – eine wesentliche Schwierigkeit entsprechender Ansätze ist demnach nicht das Erreichen maximaler Werte in Bezug auf Einzelmaße, sondern vielmehr die Entwicklung generischer Musterkataloge mit stabiler Performanz auf unterschiedlichen Datensätzen, d. h. es sind Muster zu definieren, die bei möglichst maximaler

Präzision eine möglichst maximale Abdeckung erreichen. Für ein solches Ziel wird jedoch – auch hierauf war bereits mehrfach hingewiesen worden – die Entwicklung systematischer und möglichst mathematisch formalisierter Such- bzw. Lernmethoden auf umfangreichen Datensätzen erforderlich sein. Ein entsprechender Ansatz könnte beispielsweise im Erlernen und Verallgemeinern lexikalisierter Dependenzstrukturen für *knowledge patterns* bestehen und würde damit über die in der Literatur beschriebenen überwiegend sequentiell ausgerichteten Ansätze hinausgehen. Ein wesentliches Hemmnis dafür liegt allerdings darin, dass die Voraussetzung entsprechend annotierte Korpora ausreichender Größe wären.

Nicht zuletzt zeigen die beschriebenen Experimente aber auch, dass musterbasierte Ansätze – zumindest in der hier dargestellten Form der Umsetzung – eine gewisse Leistungsgrenze aufzuweisen scheinen: In Kapitel 3 war bereits auf entsprechende kritische Befunde in der Literatur verwiesen worden und auch in der Analyse der hier erreichten Extraktionsergebnisse wurden Aspekte wie idiosynkratische Formulierungen oder die Korpusgebundenheit bestimmter Muster diskutiert. Die Ergebnisse der punktuellen Optimierungsversuche für beide Sprachen machen zudem deutlich, dass auf Grundlage einer niedrigen Performanz bereits mit einfachen Mitteln zwar relativ deutliche Verbesserungen erreicht werden können (Sub-TKR), dies aber auf Datensätzen, auf denen bereits ein gutes Gesamtergebnis erzielt wird, eher schwer ist (Sub-EKR). Performanzwerte, wie sie für einfachere linguistische Analyseaufgaben wie beispielsweise die Wortformenanalyse oder die Wortartenannotation mittlerweile gängig sind, sind folglich auf Grundlage musterbasierter Extraktionsverfahren eher unwahrscheinlich, wobei allerdings auch ML-basierte oder hybride Methoden bislang keine vergleichbaren Ergebnisse geliefert haben. Aus terminographischer Sicht bedeutet dies, dass Extraktionsergebnisse stets durch Terminologen zu überprüfen sind, ihrer breiten praktischen Anwendung aber wahrscheinlich noch einige Hindernisse entgegenstehen (vgl. hierzu Abschnitt 6.4). Allerdings sind zumindest die implementierten Filter- bzw. Ausnahmeregeln mit den im vorigen Kapitel vorgetragenen Befunden der linguistischen Analyse im Wesentlichen konsistent. Auch ein großer Teil derjenigen *knowledge patterns*, die sich als Grundlage für Extraktionsmuster mit stabiler Performanz erwiesen haben, findet sich bereits in den in Kapitel 5 (Abschnitt 5.2.2.3) zusammengestellten Boostwort-Listen wieder. Die Ergebnisse linguistischer Analysen können mithin in die Entwicklung musterbasierter Extraktionssysteme einfließen.

Wird nun danach gefragt, ob sich aus den Besonderheiten des Russischen als einer Sprache mit stark ausgeprägten Flexionsparadigmen und freier Wortfolge besondere Schwierigkeiten für die KRC-Extraktion ergeben, zeigt sich ein gemischtes Bild. In Kapitel 3 war zwar auf die Arbeit von Przepiórkowski et al. (2007) verwiesen worden, die mit ihren musterbasierten Extraktionsmodulen für eine Reihe slawischer Sprachen insgesamt recht niedrige Performanzwerte erreichen und diesen Umstand auf Besonderheiten slawischer Sprachen zurückführen: Insbesondere führten morphologische Ambiguitäten zu fehlerhaften Annotations- und Lemmatisierungsergebnissen, welche die Extraktion erschwerten, wohingegen die Stellungsfreiheit der Satzglieder durch umfangreichere Regelsätze abgefangen werden müsse. Allerdings weist auch das Deutsche – wenn auch in schwächerer Ausprägung als das Russische – die genannten Besonderheiten auf. Im Ergebnis der durchgeführten Experimente erscheint die freie Wortfolge des Russischen ebenso wie die ebenfalls nicht vollständig regulierte Wortfolge des Deutschen tatsächlich nicht als grundsätzliches Problem, da der gewählte syntaxbasierte Ansatz der Musterformulierung von konkreten Wortsequenzen abstrahiert. Zwar waren – wie am Beispiel der sequentielle Informationen nutzenden Muster *agent1* und *agent2* diskutiert – tatsächlich Schwierigkeiten im Umgang mit sequentieller Information festgestellt worden, jedoch bezog sich diese Problematik nur auf einen geringen Teil sowohl der deutschen als auch der russischen Muster. Auch inkonsistente oder fehlerhafte

Annotationsergebnisse auf den russischen Korpora wurden diskutiert und als Fehlerquelle insbesondere in der Phase der Term-Erkennung ausgemacht, als Hauptfehlerursache für falsche Negative wurden allerdings fehlende Extraktionsmuster genannt[134]. Fehlerhafte oder zumindest inkonsistente Annotationen spielen mithin zwar eine die Extraktion insgesamt erschwerende Rolle, jedoch erweist sich die Extraktion wissenshaltiger Kontexte aus russischen Korpora als grundsätzlich schwieriges Unterfangen, und zwar aus den genannten linguistischen Gründen, nämlich dem Umstand, dass das Sub-TKR eine Vielzahl von verblosen Bindestrich-KRCs enthält, die mit vertretbarer Präzision nur schwer zu extrahieren sind, und, wie dargestellt, ein hohes Maß an linguistischer Variabilität aufweist. Da entsprechende Schwierigkeiten indes auch im Umgang mit anderen Sprachen oder Korpora auftreten können, lässt sich aus diesem Befund jedoch keine Sonderstellung des Russischen oder slawischer Sprachen allgemein ableiten. Im Vergleich zum Deutschen erweist sich die eingeschränkte Verfügbarkeit von Datensätzen, Ressourcen und Analysewerkzeugen damit zwar als grundsätzliches Problem der Arbeit mit dem Russischen, zumal die für wissenschaftliche Zwecke frei verfügbaren Ressourcen und Werkzeuge nicht immer eine vergleichbare Leistungsfähigkeit oder Qualität aufweisen oder aufwendig angepasst werden müssen (wie etwa das zur Satzgrenzenerkennung verwendete Werkzeug, vgl. Kapitel 4), von einer Charakterisierung des Russischen als grundsätzlich „schwieriger“ oder „schwieriger zu bearbeiten“ soll hier jedoch mit aller Vorsicht, die angesichts dieser vorläufigen Einschätzung geboten ist, Abstand genommen werden.

Damit kann – zumindest vorläufig – auch die Frage beantwortet werden, ob ein syntaxbasierter Ansatz für das Russische grundsätzlich geeignet ist: Dies scheint – wenn einmal von der Abhängigkeit eines solchen Ansatzes von der Verfügbarkeit entsprechender Analysewerkzeuge mit einer ausreichenden Leistungsfähigkeit abgesehen wird – in der Tat insofern der Fall zu sein, als auch für das Russische (zum Vergleich werden die in Kapitel 3 referierten Arbeiten zu slawischen Sprachen herangezogen) kompetitive Extraktionsergebnisse erzielt wurden. Zwar wurden keine Vergleichsexperimente auf Grundlage sequentieller Suchmuster durchgeführt. Jedoch kann als grundsätzliches Argument für die Eignung eines syntaxbasierten Ansatzes der Umstand angeführt werden, dass die *Meaning Text Theory* (Mel’čuk et al. 1995), auf der die genutzten syntaktischen Dependenzstrukturen ja beruhen, auch für die Beschreibung des Russischen entwickelt wurde. Des Weiteren zeigt sich, dass Extraktionsmuster, die sequentielle Informationen nutzen – von den restriktiv formulierten Mustern *tire1* und *tire4* einmal abgesehen – stark schwankende oder insgesamt niedrige Präzisionswerte aufweisen. Dieses Ergebnis spiegelt auch die Resultate der individuellen Performanzanalyse für die deutschen Extraktionsmuster.

6.3 Maschinelles Lernen

In diesem Abschnitt soll einer weiteren der einleitend aufgeworfenen Fragen nachgegangen werden, und zwar derjenigen nach der Nützlichkeit der im letzten Kapitel beschriebenen (korpus-)linguistischen Merkmale für die Extraktion wissenshaltiger Kontexte, wobei die im letzten Abschnitt entwickelten und evaluierten Extraktionsmuster ausdrücklich aus dieser Gruppe von

[134] Inwieweit dieser Umstand nicht nur auf die beschriebenen Grenzen musterbasierter Ansätze überhaupt, sondern auch auf intrinsische Ressourcenbeschränkungen oder aber Limitationen (geringe Größe, fachlich-textuelle Spezifik) des verwendeten Entwicklungskorpus‘ zurückzuführen ist, kann an dieser Stelle nicht mit letzter Sicherheit festgestellt werden.

Merkmalen ausgenommen werden sollen. Konkret geht es also um die Nützlichkeit quantitativer linguistischer Merkmale für die Unterscheidung zwischen KRCs und generischen Sätzen.
Umgesetzt und evaluiert wird die Untersuchung dieser Fragestellung mit Hilfe von Methoden des maschinellen Lernens. Für eine solche Vorgehensweise spricht die Tatsache, dass in der relevanten wissenschaftlichen Literatur ML-Verfahren den symbolischen, musterbasierten Ansätzen häufig als überlegen gegenübergestellt oder aber mit diesen zu hybriden Ansätzen kombiniert werden – die Nutzung eines entsprechenden Ansatzes erlaubt mithin nicht nur den Anschluss an den aktuellen Forschungsstand, sondern auch eine Untersuchung der Frage, wie musterbasierte und ML-Methoden zweckdienlich miteinander verbunden werden können. Aufgrund ihrer im Vergleich zu musterbasierten Verfahren grundsätzlich anderen Funktionsweise versprechen zudem ML-Methoden – je nach gewähltem Algorithmus – wenigstens in Ansätzen die Vermeidung bzw. Überwindung derjenigen Defizite, die dem symbolischen Ansatz zugeschrieben wurden (Formulierung zu weniger oder zu eng gefasster deterministischer Regeln ohne Berücksichtigung probabilistischer Elemente, inhärente Leistungsgrenze insbesondere aufgrund der eingeschränkten Abdeckung, keine Gewichtung linguistischer Merkmale nach deren Relevanz für die konkret gestellte Extraktionsaufgabe)[135]. Die am Ende des letzten Kapitels durchgeführte multivariate Analyse linguistischer Merkmale von KRCs deutete zudem darauf hin, dass bereits mittels einer – eher einfachen – logistischen Regression eine, wenngleich recht grobe, Separation der Klassen KRC vs. generisch erreicht werden kann. Die Nützlichkeit dieses Ansatzes soll nun erprobt werden. Zur Anwendung kommt dabei das bereits eingeführte Verfahren der logistischen Regression, für das in R (R Core Development Team 2012) leicht verwendbare Implementierungen existieren und das zudem – im Gegensatz zu anderen ML-Algorithmen – einer direkten linguistischen Interpretation zugänglich ist.

6.3.1 Russisch

6.3.1.1 Verringerung des Suchraums durch Vorsortierung der Korpussätze

6.3.1.1.1 Grundsätzliche Vorgehensweise

Musterbasierte Extraktionsverfahren arbeiten, wie dargestellt, in deterministischer Art und Weise auf der Gesamtheit desjenigen Korpus‘, aus dem KRC-Kandidaten extrahiert werden sollen: Sobald ein bestimmtes Muster einen Satz des Korpus‘ *matcht*, wird dieser extrahiert. Die Gesamtmenge der Sätze eines Korpus‘ konstituiert mithin den Suchraum, innerhalb dessen mit Hilfe der durch die Extraktionsmuster ausgedrückten Hypothesen nach entsprechenden *matches* gesucht wird. Ausgehend von dieser Überlegung lässt sich nun die Hypothese aufstellen, dass durch eine geeignete Vorsortierung der Korpussätze der Suchraum auf geeignete Weise verkleinert und die Präzision eines (auf diesen verkleinerten Suchraum angewandten) musterbasierten Verfahrens entsprechend verbessert werden kann. Dafür soll nun konkret auf dem Entwicklungskorpus – im Falle des Russischen das Sub-EKR – auf Grundlage der im letzten Kapitel erarbeiteten KRC-typischen linguistischen Merkmale ein Regressionsmodell trainiert werden. Mit Hilfe dieses Modells sollen

[135] Freilich darf an dieser Stelle nicht verschwiegen werden, dass Methoden des maschinellen Lernens ebenso wie musterbasierte Verfahren von zwei Faktoren ganz wesentlich abhängen, und zwar der Qualität der vorhergehenden linguistischen Analyse einerseits, die sich in einer möglichst exakten Extraktion linguistischer Merkmale niederschlagen sollte, sowie der Qualität des verwendeten Trainingsmaterials andererseits, d. h. des Umfangs und der Aussagekraft des verwendeten Trainingskorpus‘.

alsdann die Sätze des Sub-TKR hinsichtlich der Wahrscheinlichkeit, dass sie einen KRC konstituieren, bewertet werden. Wird für die durch das Regressionsmodell für jeden Satz x des Sub-TKR ausgegebene Wahrscheinlichkeit $\pi(x) = P(Y = 1|X = x)$ ein passender Schwellenwert festgelegt, auf dessen Grundlage eine binäre Klassifikationsentscheidung KRC-Kandidat vs. generischer Satz getroffen werden kann, können Korpussätze mit einer unterhalb des gewählten Schwellenwerts liegenden Wahrscheinlichkeit $\pi(x)$ auf dieser Grundlage aus dem Korpus entfernt werden. Auf das so verkleinerte Korpus wird alsdann das musterbasierte Extraktionsverfahren angewandt.

6.3.1.1.2 Daten und Modellierung

Das Regressionsmodell wird wie bereits im letzten Kapitel auf einem balancierten , d. h. einem zu gleichen Teilen aus KRCs und generischen Sätzen bestehenden Trainingsdatensatz errechnet, wobei nun allerdings – da nicht mehr linguistische Merkmale von KRCs in unterschiedlichen Korpora, sondern insbesondere die Übertragbarkeit der Methode von einem Korpus auf ein anderes getestet werden soll – keine Vermischung des Sub-EKR mit dem Sub-TKR mehr stattfindet. Der Trainingsdatensatz umfasst mithin die 293 KRCs, die als theoretisch extrahierbare KRCs im Sub-EKR identifiziert wurden, sowie 293 zufällig ausgewählte generische Sätze aus demselben Korpus. Für diese Sätze können nun in Anlehnung an die Ausführungen des vorigen Kapitels Vektoren für die in Tabelle 75 zusammengefassten linguistischen Merkmale erstellt werden.

Merkmalstyp	**Ausprägung**	**Datentyp**
Allgemein	Satz-ID	numerisch
	Satzlänge	numerisch
Zielvariable	KRC	binär
Boost- und Stopwörter, typische Wortarten	Anzahl der Nomina (*tag* N)	numerisch
	Anteil der Nomina	numerisch
	Anzahl der Adjektive (*tag* A)	numerisch
	Anteil der Adjektive	numerisch
	Anzahl der Präpositionen (*tag* S)	numerisch
	Anteil der Präpositionen	numerisch
	Anzahl der Verben (*tag* V)	numerisch
	Anteil der Verben	numerisch
	Anzahl der Konjunktionen (*tag* C)	numerisch
	Anteil der Konjunktionen	numerisch
	Anzahl der Pronomen (*tag* P)	numerisch
	Anteil der Pronomen	numerisch
	Anzahl der Adverbien (*tag* R)	numerisch
	Anteil der Adverbien	numerisch
	Anzahl der Partikel (*tag* Q)	numerisch
	Anteil der Partikel	numerisch
	Anzahl von Boostwörtern (Lemmata *вид*, *совокупность*, *из*, *который*, *непосредственно*, *и*, *себя*)	numerisch
	Anteil von Boostwörtern	numerisch
	Anzahl von Stopwörtern (Lemmata *новый*, *даже*, *очень*, *если*, *необходимость*, *не*, *нет*, *после*, *вы*, *ее*, *как*, *так*, *быть*, *должен*, *мочь*, *неоходимо*, *почти*, *мало*, *многий*, *сложный*, *следовать*, *чем*, *кроме*, *поэтому*, *этот*, *следует*)	numerisch
	Anteil von Stopwörtern	numerisch

	Anteil Boostwörter minus Anteil Stopwörter	numerisch
Grammatische und strukturelle Merkmale	Anzahl von Verben mit KRC-typischer Morphologie (imperfektiver Aspekt, Passiv, dritte Person, Tempus Präsens, Singular, Plural)	numerisch
	Anteil von Verben mit KRC-typischer Morphologie	numerisch
	Anzahl von Verben mit KRC-untypischer Morphologie (perfektiver Aspekt, Imperativ, Infinitiv, Partizip, erste Person, zweite Person, Futur, Vergangenheit)	numerisch
	Anteil von Verben mit KRC-untypischer Morphologie	numerisch
	Anteil von Verben mit KRC-typischer Morphologie minus Anteil von Verben mit KRC-untypischer Morphologie	numerisch
Terminologische Merkmale	Anzahl terminologischer Nomina	numerisch
	Term in Subjektposition	binär
	Term als präpositionales Objekt	binär
	Term als direktes Objekt	binär
	Term als Genitivattribut	binär
	Term als Instrumentalobjekt	binär
	Term in satzinitialer Position (erstes, zweites oder drittes *token*)	binär
Phraseologie	KRC-typische Phrase (*в зависимости*, *при помощи*, *с использованием*, *на основании*, *с целью*, *в случае*, *в большинстве*, *По способу*, *для обеспечения*, *для изготовления*, *Для создания*, *по роду*, *по способу*, *для выполнения*, *По виду*, *по величие*, *По назначению*, *по форме*, *под действием*, *по типу*, *в качестве*)	binär

Tabelle 75: Merkmale für die Klassifikation russischer KRC-Kandidaten

Tabelle 76 gibt ein Beispiel der Merkmalsannotation für den generischen Satz: „Дефекты на зеркале цилиндра устраните притиркой или шлифовкой.“ („Entfernen Sie Defekte auf dem Zylinderspiegel durch Schmirgeln oder Schleifen.“).

Merkmal	**Erklärung**	**Wert**
ID	Satz-ID	6028
SENT_L	Satzlänge	9
KRC	KRC	0
NOUNS_N	Anzahl der Nomina	5
NOUNS_PROP	Anteil der Nomina	0,6
ADJA_N	Anzahl der Adjektive	0
ADJA_PROP	Anteil der Adjektive	0
S_N	Anzahl der Präpositionen	1
S_PROP	Anteil der Präpositionen	0,1
V_N	Anzahl der Verben	1
V_PROP	Anteil der Verben	0,1
C_N	Anzahl der Konjunktionen	1
C_PROP	Anteil der Konjunktionen	0,1
P_N	Anzahl der Pronomen	0
P_PROP	Anteil der Pronomen	0
R_N	Anzahl der Adverbien	0
R_PROP	Anteil der Adverbien	0
Q_N	Anzahl der Partikel	0
Q_PROP	Anteil der Partikel	0

BOOST_N	Anzahl von Boostwörtern	0
BOOST_PROP	Anteil von Boostwörtern	0
STOP_N	Anzahl von Stopwörtern	0
STOP_PROP	Anteil von Stopwörtern	0
BOOST_STOP_R	BOOST_PROP-STOP_PROP	0
BOOSTM_N	Anzahl von Verben mit KRC-typischer Morphologie	1
BOOSTM_PROP	Anteil von Verben mit KRC-typischer Morphologie	0,1
STOPM_N	Anzahl von Verben mit KRC-untypischer Morphologie	1
STOPM_PROP	Anteil von Verben mit KRC-untypischer Morphologie	0,1
BOOSTM_STOPM_PROP	BOOSTM_PROP-STOPM_PROP	0
TERM_N	Anzahl terminologischer Nomina	0
TERM_S	Term in Subjektposition	0
TERM_PO	Term als präpositionales Objekt	0
TERM_DO	Term als direktes Objekt	0
TERM_GA	Term als Genitivattribut	0
TERM_IO	Term als Instrumentalobjekt	0
TERM_INI	Term in satzinitialer Position (erstes, zweites oder drittes *token*)	0
PHRASE	KRC-typische Phrase	0

Tabelle 76: Merkmalsvektor für einen russischen Beispielsatz

Bei der Berechnung des Regressionsmodells wurde derselben Methodologie gefolgt wie in Kapitel 5, d. h. absolute Frequenzen und solche Merkmale, für die eine deutliche Korrelation mit anderen Merkmalen festgestellt werden konnte, wurden in die Modellierung nicht mit einbezogen. Das von R errechnete Ausgangsmodell erreicht eine Devianz von 367,23 im Vergleich zu einer Nulldevianz von 812,37. Ein Signifikanztest auf Grundlage der Devianzdifferenzen für die einzelnen Parameter des Modells gemäß Sachs/Hedderich (2009[13]: 679 f.) ergibt, dass die folgenden Variablen signifikant (auf einem Niveau von 0,05) zur Minderung der Devianz im Vergleich zum Nullmodell beitragen:

- SENT_L,
- ADJA_PROP,
- S_PROP,
- V_PROP,
- P_PROP,
- R_PROP,
- Q_PROP,
- BOOST_STOP_R,
- BOOSTM_STOPM_PROP,
- TERM_N,
- TERM_S,
- TERM_INI,
- PHRASE.

Auch dieses Modell wurde wieder einem Prozess der Merkmalsauswahl auf Grundlage des AIC-Kriteriums (ebd.: 685 f.) unterworfen. Das so entstehende Modell selektiert die folgenden Variablen,

von denen bis auf C_PROP alle gemäß dem bereits mehrfach angewandten Test auf Grundlage der Devianzdifferenzen signifikant zur Verminderung der Modelldevianz beitragen:

- SENT_L,
- ADJA_PROP,
- S_PROP,
- C_PROP,
- P_PROP,
- Q_PROP,
- BOOST_STOP_R,
- BOOSTM_STOPM_PROP,
- TERM_N,
- TERM_S,
- TERM_INI,
- PHRASE.

Dieses Modell erreicht eine Devianz von 369,64, die Tests, die bereits in Kapitel 5 zur Evaluation der Modellanpassung durchgeführt wurden, ergeben keine Hinweise auf eine schlechte Modellanpassung bzw. eine Verletzung der der Modellierung zugrundeliegenden Annahmen.

6.3.1.1.3 Ergebnisse

Mit Hilfe des errechneten Regressionsmodells wurden die 5674 Sätze des Sub-TKR hinsichtlich ihres KRC-Status‘ bewertet, wobei die in den vorhergehenden Abschnitten erstellte Korpusvariante mit verbesserter Termerkennung herangezogen wurde. Dieses Korpus enthält 267 theoretisch extrahierbare Goldstandard-KRCs. Abbildung 33 visualisiert die sich aus dem Ranking der Korpussätze nach den von dem Regressionsmodell ausgegebenen Wahrscheinlichkeiten ergebende Präzisions-Abdeckungs-Kurve, die wiederum Rankings nach der absoluten bzw. relativen Anzahl der Termini pro Satz[136] ebenso wie den Präzisions-Abdeckungs-Werten der musterbasierten Extraktion (es wird das beste System nach Abschnitt 6.2.3.2.4 zugrunde gelegt) gegenübergestellt wird. Die rote Linie stellt das für die einzelnen Abdeckungsniveaus jeweils erreichte f-Maß des Rankings mittels logistischer Regression dar, wenn die den Abdeckungsniveaus entsprechenden Wahrscheinlichkeiten des Regressionsmodells als Schwellenwerte für Klassifikationsentscheidungen angesehen werden, die violette Linie zeigt die f-Maße für die musterbasierte Extraktion.

Augenfällig ist zunächst, dass bereits die hier durchgeführte einfache logistische Modellierung anhand recht generischer linguistischer Merkmale im Vergleich zur musterbasierten Extraktion konkurrenzfähige Performanzwerte erreicht, dabei jedoch – eine entsprechende Wahl des Schwellenwerts für Klassifikationsentscheidungen vorausgesetzt – den Vorteil einer besseren Abdeckung bei stabilem f-Maß bietet. Im Vergleich zu Abbildung 26 (Kapitel 5) wird außerdem deutlich, dass für den hier untersuchten Datensatz tendenziell höhere Präzisionswerte erreicht werden als im Fall der multivariaten Analyse deutscher KRC-Kandidaten[137]. Auch hier erweist sich zudem die logistische Regression als den ihr gegenübergestellten einfachen *baselines* überlegen. Für das hier verfolgte Interesse einer Vorauswahl von KRC-Kandidaten für einen auf dem bereinigten

[136] Diese beiden Kurven beruhen, wie bereits in Kapitel 5, auf gruppenbezogenen Präzisionswerten.

[137] Die im Vergleich zu den deutschen Daten ähnliche Klassifikationsgenauigkeit bei höherer Präzision legt den Schluss nahe, dass dieser Umstand auch auf den höheren Anteil von KRCs im Sub-TKR zurückzuführen ist.

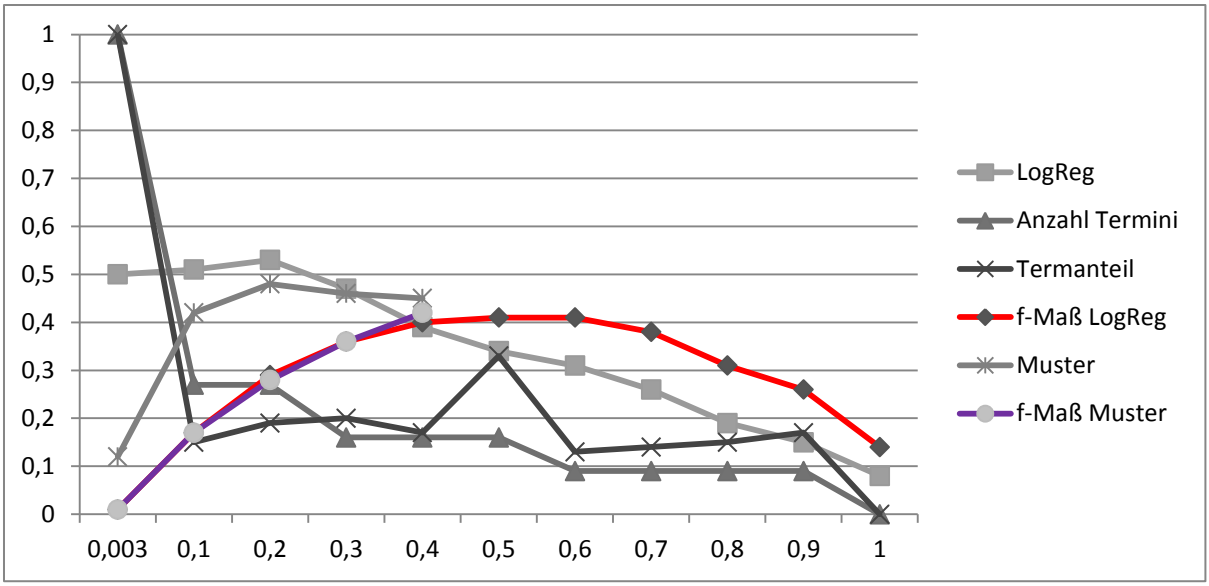


Abbildung 33: Präzisions-Abdeckungs-Kurve für das Ranking russischer KRC-Kandidaten des Sub-TKR

Korpus ansetzenden musterbasierten KRC-Extraktionsschritt relevant ist zudem der Umstand, dass 60% der aus dem Korpus extrahierbaren KRCs eine Wahrscheinlichkeit $\pi(x)$ von $\pi(x) \geq 0{,}81$ aufweisen. Den unterhalb dieses Wertes verbleibenden 5164 Sätzen wurden mithin niedrigere Werte zugeordnet. Wird wiederum der Wert 0,81 als Grenz- bzw. Schwellenwert für die Klassifikation von KRC-Kandidaten gesetzt, ergibt sich die folgende Klassifikationsgenauigkeit (*accuracy a*):

$$a = \frac{Anzahl\ der\ korrekt\ klassifizierten\ Sätze}{Anzahl\ aller\ klassifizierten\ Sätze}$$

$$= \frac{160\ KRCs + 5164\ niedriger\ gerankte\ Sätze - 107\ falsch\ klassifizierte\ KRCs}{5674} = \frac{5217}{5674} \approx 92\%$$

Dies entspricht einer Präzision von 31% und einer Abdeckung von 60%, so dass sich ein f-Maß von 0,41 ergibt. Für die Untersuchung der Frage, ob sich auf Grundlage einer solchen Klassifikation bessere musterbasierte Extraktionsergebnisse erzielen lassen, wurde eine Reihe unterschiedlicher Wahrscheinlichkeiten als Schwellenwert für die klassifikationsbasierte Auswahl von Korpussätzen gewählt. Diejenigen Sätze, denen durch den Regressionsalgorithmus eine über dem jeweiligen Schwellenwert liegende Wahrscheinlichkeit zugeordnet wurde, wurden anhand ihrer Identifikationsnummern im Korpus erkannt und einem jeweils kleineren Subkorpus hinzugefügt. Alle verbleibenden Sätze wurden für den nachfolgenden Extraktionsschritt dagegen nicht berücksichtigt. Die Extraktion wurde zudem in zwei *settings* durchgeführt, und zwar einem *setting*, in dem die musterbasierte Extraktion einschließlich der gemeinsam mit den Suchmustern implementierten Filter- bzw. Ausnahmeregeln durchgeführt wurde, und einem *setting* ohne solche Ausnahmeregeln. Tabelle 77 fasst die Ergebnisse für die unterschiedlichen *settings* und Schwellenwerte zusammen. Die Zahlen in der linken Spalte bezeichnen jeweils die gewählten Schwellenwerte. Des Weiteren wird die für jeden Schwellenwert für die musterbasierte Extraktion ausgewählte Anzahl von Korpussätzen angegeben. Die so erreichten Evaluationsmaße Präzision, Abdeckung und f-Maß werden zudem den durch eine rein musterbasierte Extraktion ermittelten Werten (vgl. die vorigen Abschnitte) sowie dem durch bloße Klassifikation höchsten erreichten f-Maß gegenübergestellt. Der in den beiden *settings* für die drei Evaluationsmaße jeweils höchste erreichte Wert ist fett markiert.

Setting	Sätze im Subkorpus	Präzision	Abdeckung	f-Maß
0,3069	1657	0,5102	**0,3745**	**0,4320**
0,4154	1355	0,5429	0,3558	0,4299
0,5140	1121	0,559	0,3371	0,4206
0,6003	932	0,5723	0,3258	0,4153
0,7119	709	0,6058	0,3108	0,4109
0,8117	508	0,6435	0,2771	0,3874
0,9171	279	**0,7125**	0,2135	0,3285
0,3069_ohneFilter	1657	0,4739	0,4082	0,4386
0,4154_ohneFilter	1355	0,5099	0,3858	**0,4392**
0,5140_ohneFilter	1121	0,5246	0,3596	0,4267
0,6003_ohneFilter	932	0,5349	0,3446	0,4191
0,7119_ohneFilter	709	0,5649	0,3258	0,4133
0,8117_ohneFilter	508	0,6141	0,2921	0,3959
0,9171_ohneFilter	279	**0,7011**	0,2285	0,3446
musterbasiert	5674	0,4518	0,3858	0,4162
musterbasiert_ohnefilter	5674	0,3979	0,4232	0,4102
LogReg	5674	0,3150	**0,5993**	0,4129

Tabelle 77: Ergebnisse der musterbasierten Extraktion nach Klassifikation der Korpussätze des Sub-TKR mit logistischer Regression in verschiedenen settings

Die Tabelle zeigt zunächst im Zusammenspiel mit Abbildung 33, dass bereits durch die einfache logistische Regression (*setting* LogReg) bei Festlegung eines optimalen Schwellenwerts ungeachtet der Generizität der linguistischen Merkmale, die der Algorithmus nutzt, für die Klassifikation Performanzwerte erreicht werden, die im Bereich der durch die manuell spezifizierten Extraktionsmuster auf dem Sub-TKR gewährleisteten Performanz liegen. Insbesondere wird durch die Klassifizierung mit logistischer Regression die beste Abdeckung (ca. 60%) erzielt. Bei der Vorselektion von Korpussätzen für die musterbasierte Extraktion werden dagegen bei der Wahl eines ausreichend hohen Schwellenwerts Präzisionswerte über 70% (*setting* 0,9171) erreicht, wohingegen selbst bei Wahl eines sehr niedrigen Schwellenwerts von 0,3069 (*setting* 0,3069) über 4000 Sätze des Sub-TKR verworfen, d. h. nicht in die Extraktion einbezogen werden. Tatsächlich ist die Wahl des Schwellenwerts für die Erlangung von Performanzsteigerungen gegenüber der musterbasierten Extraktion maßgeblich – optimale Werte, mit deren Hilfe in der überwiegenden Zahl der Fälle auf unterschiedlichen Datensätzen ein möglichst hohes f-Maß erreicht wird, können – soweit möglich – jedoch nur auf Grundlage umfangreicher Testreihen festgelegt werden. Insgesamt ist die durch eine optimale Wahl der Parameter erreichte Performanzsteigerung in diesem Szenario jedoch gering – sie beträgt, bezogen auf das f-Maß, 0,4392-0,4162=0,0230 Punkte. Es besteht Grund zu der Annahme, dass diese zwar bemerkenswerte, aber doch eher unbefriedigende Verbesserung auch durch die inhärenten Grenzen der im zweiten Extraktionsschritt angewandten Extraktionsmuster begrenzt wird, d. h. insbesondere das Problem der recht geringen Abdeckung aufgrund fehlender oder zu restriktiv formulierter Muster kann auf diese Weise nicht überwunden werden. Diese Interpretation der Ergebnisse wird auch durch die Beobachtung gestützt, dass in dem *setting* ohne einschränkende Filter bzw. Ausnahmeregeln stets ein höheres f-Maß erzielt wird als in dem *setting* mit Filtern.

6.3.1.2 Einbezug der Musterinformation in die logistische Modellierung

6.3.1.2.1 Grundsätzliche Vorgehensweise

Im vorigen Abschnitt wurde die Hypothese formuliert, dass die nachfolgende Verwendung von Extraktionsmustern im Anschluss an ein mathematisches Ranking- bzw. Klassifikationsverfahren grundlegende Probleme musterbasierter Ansätze in die KRC-Extraktion einbringt, so dass das Verbesserungspotenzial des ML-Ansatzes nicht vollständig genutzt werden kann. Eine Lösung für dieses Problem könnte nun darin bestehen, dass musterbezogene Informationen – z. B. die Information darüber, ob ein Satz ein bestimmtes oder mehrere Extraktionsmuster *matcht* und wie präzise dieses Muster bisher war – nicht erst nach einer abgeschlossenen Sortierung oder Vorauswahl von Korpussätzen beispielsweise durch logistische Regression zum Einsatz kommen, sondern in diese einfließen. In diesem Fall bleibt die Gewichtung der einzelnen Klassifikationsmerkmale einschließlich der musterbezogenen Informationen dem Algorithmus überlassen, so dass beispielsweise die Möglichkeit besteht, *matches* unpräziser Extraktionsmuster niedriger zu bewerten als solche sehr zuverlässiger Muster. Dieser Ansatz soll nun erprobt werden.

6.3.1.2.2 Daten und Modellierung

Die Trainingsdaten für die erneute logistische Modellierung umfassen wieder die 293 aus dem Sub-EKR extrahierbaren KRCs sowie 293 zufällig ausgewählte generische Sätze aus demselben Korpus. Ein Teil der bisher verwendeten linguistischen Merkmale für die Beschreibung der Daten wird jedoch nun zu zwei größeren Kategorien zusammengefasst, und zwar:

- GUT: Diese Kategorie umfasst die Wortarten Nomen, Adjektiv, Präpositionen und Pronomen. Diesen Wortarten wurden durch die in den vorigen Abschnitten beschriebenen Modelle positive Regressionskoeffizienten zugeordnet, es handelt sich also um für KRCs eher typische Wortarten. Aufgrund des positiven Regressionskoeffizienten werden auch die Pronomen, die im letzten Kapitel den Negativmerkmalen zugeordnet wurden, in diese Kategorie einbezogen.
- SCHLECHT: Diese Kategorie umfasst die Wortarten Verb, Konjunktion, Adverb und Partikel.

Des Weiteren wird eine Reihe zusätzlicher Merkmale modelliert, und zwar die Information darüber, wie viele der manuell spezifizierten Extraktionsmuster einen Satz extrahieren, Informationen über die Präzision dieser Muster, Informationen darüber, ob in der näheren Umgebung eines Satzes noch weitere Sätze durch Extraktionsmuster *gematcht* werden, Informationen darüber, ob der Satz mit einer Nominalphrase beginnt, Informationen über den Abstand zwischen einem Terminus in Subjektposition und dem zugehörigen Prädikat und Informationen über das letzte *token* in einem Satz. Tabelle 78 gibt einen Überblick über den neuen Merkmalskatalog.

Merkmalstyp	**Ausprägung**	**Datentyp**
Allgemein	ID	numerisch
	SENT_L	numerisch
Zielvariable	KRC	binär
Extraktionsinformation	PATTERN_N: Anzahl der Muster, die einen Satz extrahieren	numerisch

	PATTERN_V: Mittelwert aus den Präzisionswerten für die Muster aus PATTERN_N	numerisch
	NEAR_HITS: musterbasierte Treffer in der näheren Umgebung eines Satzes	binär
Boost- und Stopwörter, typische Wortarten	GUT	numerisch
	GUT_PROP: Anteil „guter“ Wortarten im Satz	numerisch
	SCHLECHT	numerisch
	SCHLECHT_PROP: Anteil „schlechter“ Wortarten im Satz	numerisch
	BOOST_N	numerisch
	BOOST_PROP	numerisch
	STOP_N	numerisch
	STOP_PROP	numerisch
	BOOST_STOP_R	numerisch
Grammatische und strukturelle Merkmale	FIRST_CHUNK: Satz beginnt mit einer Nominalphrase	binär
	BOOSTM_N	numerisch
	BOOSTM_PROP	numerisch
	STOPM_N	numerisch
	STOPM_PROP	numerisch
	BOOSTM_STOPM_PROP	numerisch
	ABSTAND: Abstand zwischen Subjektterm und Prädikat	numerisch
	LAST: Satz ist ein Aussagesatz und endet auf einen Punkt.	binär
Terminologische Merkmale	TERM_N	numerisch
	TERM_S	binär
	TERM_PO	binär
	TERM_DO	binär
	TERM_GA	binär
	TERM_IO	binär
	TERM_INI	binär
Phraseologie	PHRASE	binär

Tabelle 78: Erweiterter Merkmalskatalog

Die Angabe für die Anzahl der Muster, die einen Korpussatz extrahieren, basiert dabei auf dem in Abschnitt 6.2.3.2.4 beschriebenen Musterkatalog und entspricht damit der besten musterbasierten Extraktionsperformanz für das Sub-TKR. Die Angaben für das Merkmal PATTERN_V sind Mittelwerte aus Präzisionsbewertungen der einzelnen Muster und reichen von 0 (=der Satz wird durch kein Muster extrahiert) über 1 (=der Satz wird durch unpräzise Muster extrahiert) und 2 (=der Satz wird durch Muster mit wechselhafter Präzision extrahiert) bis maximal 3 (=der Satz wird durch präzise Muster extrahiert). Jedem Extraktionsmuster wurde dafür ein Präzisionswert zugeordnet, der entweder 1, 2 oder 3 beträgt, wobei als Grundlage der Bewertung die in Tabelle 74 (Abschnitt 6.2.3.2.4) zusammengefasste Analyse der Musterpräzision diente. Durch die Mittelwertbildung können sich für dieses Merkmal Werte mit Dezimalstellen ergeben, so dass es in der Modellierung nicht als Faktor, sondern als numerisches *feature* kodiert wird. Das Merkmal NEAR_HITS gibt an, ob die zu beiden Seiten eines Korpussatzes nächstliegenden fünf Sätze ebenfalls durch

Extraktionsmuster *gematcht* werden, und soll dem in der Literatur genannten Umstand Rechnung tragen, dass KRCs in Gruppen auftreten. Das Merkmal FIRST_CHUNK kodiert die mittels einer einfachen Heuristik extrahierte Information darüber, ob ein Korpussatz mit einer Nominalphrase beginnt. Das Merkmal ABSTAND misst, wie nah ein Term in Subjektposition an seinem Prädikat gelegen ist und trägt somit zur Beschreibung syntaktischer Strukturen bei. Gibt es keinen Terminus in Subjektposition, beträgt der Wert des Merkmals 0. Schließlich gibt das Merkmal LAST an, ob ein Korpussatz auf einen Punkt endet und entsprechend mit hoher Wahrscheinlichkeit ein Aussagesatz ist. Tabelle 79 stellt die erweiterte Annotation der linguistischen Merkmale wiederum an einem Beispiel dar, und zwar dem KRC: „Привод клапанов осуществляется от распределительного вала при помощи установленных непосредственно под кулачками приводных рычагов.“ („Die Ventile werden durch die Nockenwelle unter Verwendung der unmittelbar unterhalb der Nocken gelegenen Antriebshebel gesteuert.“). Die Angaben in Klammern beziehen sich auf Extraktionsmuster, Wortarten-*tags* und konkrete Werte für die untersuchten Merkmale.

Merkmal	**Wert**
ID	1201
SENT_L	15
KRC	1
PATTERN_N	1 (*осуществляться1*)
PATTERN_V	2 (*осуществляться1*)
NEAR_HITS	1 (Satz 1200)
GUT	11 (*Привод* (N)*, клапанов* (N)*, от* (S)*, распределительного* (A)*, вала* (N)*, при* (S)*, помощи* (N)*, под* (S)*, кулачками* (N)*, приводных* (A)*, рычагов* (N))
GUT_PROP	0,7
SCHLECHT	3 (*осуществляется* (V)*, установленных* (V)*, непосредственно* (R))
SCHLECHT_PROP	0,2
BOOST_N	1 (*непосредственно*)
BOOST_PROP	0,1
STOP_N	0
STOP_PROP	0
BOOST_STOP_R	0,1
FIRST_CHUNK	1
BOOSTM_N	2 (beide Verben)
BOOSTM_PROP	0,1
STOPM_N	1 (*установленных* – Partizip, Vergangenheit, perfektiv)
STOPM_PROP	0,1
BOOSTM_STOPM_PROP	0
ABSTAND	0
LAST	1
TERM_N	1 (*вала* – zu *распределительный вал*)
TERM_S	0
TERM_PO	1 (*осуществляется от распределительного вала*)
TERM_DO	0
TERM_GA	0
TERM_IO	0
TERM_INI	0
PHRASE	1 (*при помощи*)

Tabelle 79: Erweiterter Merkmalsvektor für einen Beispielsatz

Die Berechnung des Regressionsmodells erfolgte wiederum entsprechend der bereits mehrfach beschriebenen Methodologie. Das am Ende des Prozesses, d. h. nach Merkmalsselektion mit Hilfe des AIC-Kriteriums (Sachs/Hedderich (2009[13]: 685 f.) errechnete Modell selektiert die in Tabelle 80 zusammengefassten Merkmale und erreicht eine Devianz von 313,43 im Vergleich zu einer Nulldevianz von 812,37. Auch für dieses Modell ergeben die durchgeführten Prüfungen keine Hinweise auf eine schlechte Modellanpassung oder die Verletzung von Modellannahmen. Die Tabelle enthält für jedes der selektierten Merkmale neben dem Regressionskoeffizienten und dem Standardfehler auch das dazugehörige Konfidenzintervall sowie eine Angabe zur Signifikanz des jeweiligen Parameters gemäß dem Chi-Quadrat-Test auf Grundlage der Devianzdifferenzen (ebd.: 679 f.).

Merkmal	Koeffizient	Standardfehler	Konfidenzintervall	Signifikanz
PATTERN_V	0,9744	0,2238	[0,5754, 1,4677]	ja
NEAR_HITS1	0,9143	0,2934	[0,3428, 1,4963]	ja
SENT_L	0,0516	0,0136	[0,0258, 0,0796]	ja
GUT_PROP	5,4768	1,2431	[3,1197, 8,0077]	ja
BOOST_STOP_R	7,0497	2,4683	[2,2876, 11,9851]	ja
BOOSTM_STOPM_PROP	10,0247	2,4876	[5,2920, 15,0674]	ja
TERM_N	0,3460	0,1499	[0,0676, 0,6561]	ja
TERM_S1	0,9306	0,3974	[0,1506, 1,7140]	ja
TERM_GA1	-1,1156	0,3960	[-1,9190, -0,3621]	ja
TERM_INI1	1,0923	0,3681	[0,3763, 1,8246]	ja
LAST1	3,2457	0,8852	[1,7632, 5,3849]	ja

Tabelle 80: Modellparameter für ein u. a. um musterbezogene Informationen erweitertes Regressionsmodell

Aus der Tabelle geht hervor, dass dieses Modell – bis auf das Merkmal TERM_GA – ausschließlich positive Merkmale für das Vorliegen eines KRCs selektiert, wobei neben Informationen zur Präzision *matchender* Suchmuster und zu benachbarten *matches* auch strukturelle, lexikalische und terminologische Informationen genutzt werden. Der negative Koeffizient des Merkmals TERM_GA deutet zudem darauf hin, dass – obwohl abweichende Konstellationen im Zuge der im letzten Kapitel durchgeführten Analyse durchaus gefunden wurden – die reguläre Position des Definiendums, d. h. die Positionierung als satzinitiales Subjekt in quantitativer Hinsicht eher als KRC-typisch anzusehen ist. Dieser Befund bestätigt auch das Ergebnis der im letzten Kapitel durchgeführten Regressionsanalyse.

6.3.1.2.3 Ergebnisse

Mit Hilfe des erweiterten Regressionsmodells wurden wiederum die 5674 Sätze des Sub-TKR hinsichtlich ihres KRC-Status‘ bewertet. Abbildung 34 visualisiert die sich aus der Sortierung der Korpussätze nach den ausgegebenen Wahrscheinlichkeiten ergebende Präzisions-Abdeckungs-Kurve (erweitertes Modell, Stepmodel2_ru), die der Kurve des einfachen Regressionsmodells aus Abschnitt 6.3.1.1.3 (Stepmodel1_ru) ebenso wie der Präzisions-Abdeckungs-Kurve für die musterbasierte Extraktion gegenübergestellt wird. Für alle Modelle sind zudem die Kurven für das f-Maß dargestellt.

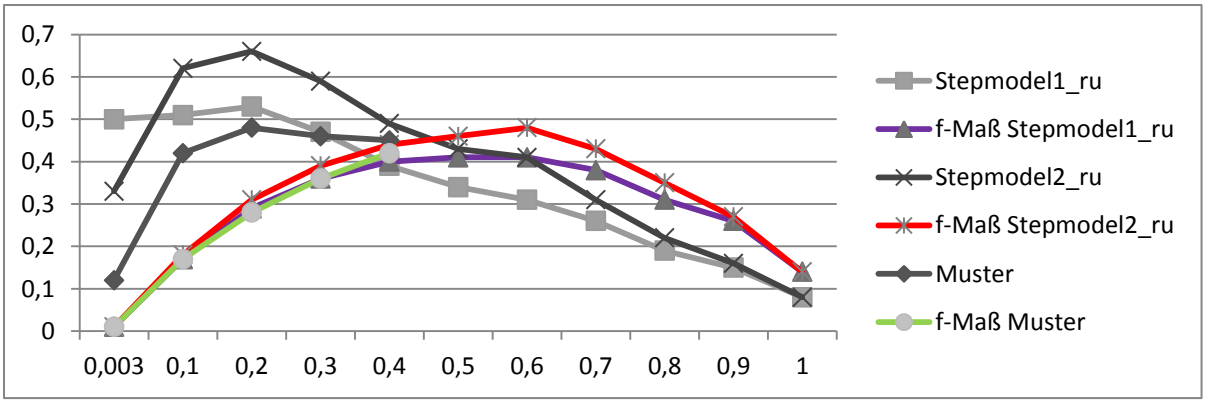


Abbildung 34: Präzisions-Abdeckungs-Kurve für das Ranking russischer KRC-Kandidaten des Sub-TKR mit dem erweiterten Regressionsmodell im Vergleich zu vorher erreichten Werten

Die Abbildung zeigt einen deutlichen und über weite Strecken konstanten Präzisionsgewinn durch das erweiterte Modell, der sich auch in einem deutlich höheren maximalen f-Maß von 0,48 niederschlägt. Dieses f-Maß entspricht einer Klassifikationsgenauigkeit von 94%. Hiermit ist im Vergleich zur musterbasierten Extraktion, aber auch zum vorhergehenden Ranking-Experiment eine relevante Verbesserung erreicht. Ein Vergleich der für das erweiterte Modell errechneten Regressionskoeffizienten mit den (dort allerdings nicht dargestellten) Koeffizienten des in Abschnitt 6.3.1.1.2 errechneten einfachen Regressionsmodells legt nahe, dass dieser Gewinn auf die Hinzunahme der extraktionsbezogenen Informationen zu Mustern und benachbarten Treffern ebenso wie auf das Merkmal LAST mit seinem relativ hohen Regressionskoeffizienten zurückzuführen ist. Die Bedeutung der terminologischen Merkmale scheint dagegen – abgesehen von der Hinzunahme des Merkmals TERM_GA mit einem recht hohen Koeffizienten für KRC-untypische Strukturen – beim Übergang zum erweiterten Regressionsmodell weitgehend gleich geblieben bzw. leicht verringert worden zu sein. Geringer als im einfachen Modell scheint jedoch die Bedeutung lexikalischer Informationen: Das Merkmal GUT_PROP, das Informationen zu allen KRC-typischen Wortarten kodiert, erhält einen Regressionskoeffizienten, der im Bereich oder sogar noch unterhalb derjenigen Koeffizienten liegt, die im einfachen Modell den einzelnen wortartenbezogenen Merkmalen zugeordnet wurden. Entsprechend sinkt beim Übergang vom einfachen zum erweiterten Modell auch der Koeffizient des Merkmals BOOST_STOP_R von rund 12 auf rund 7. Der Koeffizient für das morphologische Merkmal BOOSTM_STOPM_PROP bleibt dagegen nahezu unverändert, was auf eine stabile Bedeutung dieses Merkmals hindeutet.

Diese Interpretation kann in derselben Weise wie bereits in Abschnitt 5.3.4.3 durch systematische Variation der Merkmalsbelegungen für die einzelnen Merkmale überprüft werden. Dafür wird aus den für die Regression verwendeten Testdaten, d. h. den Sätzen des Testkorpus‘, eine Stichprobe von 100 Sätzen gezogen. Diese Stichprobe enthält 20 KRCs. Tabelle 81 fasst die Ergebnisse dieser Analyse analog zu Tabelle 64 zusammen:

Merkmal	**Merkmal**	**Mittelwert Vorhersage**	**Signifikanz**
PATTERN_V			
setting0	0,2700	0,4549	für setting0 vs. setting2, für setting1 vs. setting2
setting1	0,1000	0,4377	
setting2	1,2700	0,5842	

NEAR_HITS			
setting0	0: 18, 1: 82	0,4549	für setting0 vs. setting1, für setting1 vs. setting2
setting1	0: 100, 1: 0	0,3522	
setting2	0: 0, 1: 100	0,4692	
SENT_L			
setting0	17,95	0,4549	nicht signifikant
setting1	16,95	0,4482	
setting2	18,95	0,4617	
GUT_PROP			
setting0	0,6120	0,4549	für setting1 vs. setting2
setting1	0,5060	0,4015	
setting2	0,7319	0,5319	
BOOST_STOP_R			
setting0	0,0360	0,4549	für setting0 vs. setting2, für setting1 vs. setting2
setting1	-0,0832	0,3912	
setting2	0,1500	0,5518	
BOOSTM_STOPM_PROP			
setting0	0,0580	0,4549	signifikant für alle drei Vergleiche
setting1	-0,0646	0,3633	
setting2	0,1747	0,6008	
TERM_N			
setting0	1,2600	0,4549	nicht signifikant
setting1	0,6000	0,4216	
setting2	2,2600	0,5009	
TERM_S			
setting0	0: 67, 1: 33	0,4549	für setting0 vs. setting2, für setting1 vs. setting2
setting1	0: 100, 1: 0	0,4158	
setting2	0: 0, 1: 100	0,5391	
TERM_GA			
setting0	0: 67, 1: 33	0,4549	für setting0 vs. setting2, für setting1 vs. setting2
setting1	0: 100, 1: 0	0,5022	
setting2	0: 0, 1: 100	0,3668	
TERM_INI			
setting0	0: 56, 1: 44	0,4549	für setting1 vs. setting2
setting1	0: 100, 1: 0	0,3904	
setting2	0: 0, 1: 100	0,5408	
LAST			
setting0	0:16, 1: 84	0,4549	für setting0 vs. setting1, für setting1 vs. setting2
setting1	0: 100, 1: 0	0,1392	
setting2	0: 0, 1: 100	0,5075	

Tabelle 81: Auswirkungen der systematischen Variation von Merkmalswerten auf einer Zufallsstichprobe aus dem Sub-TKR

Die Tabelle zeigt zunächst, dass im Rahmen der durchgeführten Analyse für alle Merkmale des erweiterten Regressionsmodells bis auf SENT_L und TERM_N durch systematische Veränderung der Merkmalsbelegungen signifikant unterschiedliche Klassifikationsergebnisse erzielt wurden. Bei einem Großteil der signifikant voneinander abweichenden Klassifikationsergebnisse beruht der Unterschied allerdings auf relativ starken Veränderungen der Einzelmerkmale, so z. B. bei dem Merkmal PATTERN_V auf der Erhöhung aller Werte um 1 (*setting2*), was angesichts der Durchschnittsbildung, die diesem Merkmal zugrunde liegt, eine eher hypothetische Veränderung darstellt. Für die binären

Merkmale gehen die signifikanten Abweichungen ebenfalls auf drastische Veränderungen zurück, allerdings deuten die Ergebnisse darauf hin, dass durch eine verbesserte Performanz der Extraktionsmuster sowie eine bessere Term-Erkennung das Verbesserungspotential der hier dargestellten Methode der KRC-Extraktion noch weiter ausgeschöpft werden kann, wobei weniger die Anzahl der Termini in einem Satz relevant zu sein scheint als deren konkrete syntaktische Position. Im Fall der lexikalischen Merkmale wirken sich ebenfalls erst relativ drastische Veränderungen signifikant auf die Klassifikationsergebnisse aus, d. h. für das Merkmal GUT_PROP zeigen sich signifikante Unterschiede nur für den Vergleich von *setting1* und *setting2*, was einem Unterschied von zwei „guten“ Einheiten pro Satz entspricht. Demgegenüber zeitigen schon leichte Veränderungen des Merkmals BOOSTM_STOPM_PROP abweichende Klassifikationsergebnisse, was für die Bedeutung des Merkmals spricht. Veränderungen des Merkmals LAST schlagen sich im Unterschied zu den Tests für die anderen Merkmale in hochsignifikanten Ergebnissen ($p < 0{,}0001$) nieder, so dass das Kriterium „ist ein Aussagesatz“ für die Identifikation von KRCs ebenfalls von großer Bedeutung zu sein scheint. Insgesamt erscheinen aufgrund der hier durchgeführten Analyse – wie bereits im Hinblick auf die Regressionskoeffizienten diskutiert – terminologische (positions- bzw. syntaxbezogene), grammatische (morphologische) und musterbezogene Merkmale als die wichtigsten Informationsquellen für die logistische Analyse der untersuchten Daten, wobei – wie bereits dargestellt – eher reguläre, d. h. linguistisch unmarkierte Strukturen (z. B. Terminus in Subjektposition, Aussagesatz) bevorzugt zu werden scheinen. Lexikalische Merkmale tragen ebenfalls zur Klassifikation bei, scheinen aber von vergleichsweise geringerer Bedeutung zu sein.

6.3.1.3. Zwischenfazit

Die vorstehenden Ausführungen haben ergeben, dass durch den Einbezug der im letzten Kapitel quantitativ erfassten linguistischen Merkmale beispielsweise im Zuge einer logistischen Modellierung z. T. deutliche Steigerungen der Extraktionsperformanz erreicht werden können. Insbesondere zeigt sich, dass bereits ein einfaches logistisches Modell auf Grundlage linguistischer Merkmale ohne jeden Einbezug musterbezogener Information (Abschnitt 6.3.1.1) eine Performanz erreicht, die zumindest für die hier analysierten Daten im Bereich der Extraktionsgüte durch manuell spezifizierte Muster liegt. Auch die Vorsortierung von Korpussätzen mit Hilfe einer logistischen Regression zeitigt eine Verbesserung der Extraktionsperformanz durch Steigerung der erreichten Präzision, ist jedoch an zwei wesentliche Nachteile gebunden, und zwar:

- die Schwierigkeit der Bestimmung eines optimalen Schwellenwerts für die Klassifikation,
- die durch die nachfolgende Verwendung der manuell spezifizierten Extraktionsmuster immer noch eingeschränkte Abdeckung.

Werden dagegen Suchmuster bzw. die durch diese erzielten *matches* einschließlich einer Präzisionsbewertung in den Modellierungsprozess einbezogen, d. h. als Merkmale unter anderen behandelt, kann durch die Wahl eines geeigneten Schwellenwerts zumindest die zweite Einschränkung in Teilen überwunden werden, wobei durch die deutlich gesteigerte Präzision eine stark verbesserte Abdeckung und damit ein höheres f-Maß erreicht werden können. Tatsächlich ordnet der im vorigen Abschnitt beschriebene und evaluierte Algorithmus auch solchen KRCs relativ hohe Wahrscheinlichkeiten zu, für die keine Extraktionsmuster existieren und die entsprechend durch die symbolische Methode nicht erkannt werden, vgl. die folgenden Beispiele, die alle auf einem 60-prozentigen Abdeckungsniveau noch extrahiert werden können.

113) При окончательном контроле проверяют качество готовых сварных соединений в соответствии с требованиями соотвествующих нормативных документов.
[*Übersetzung*: Bei der Endkontrolle wird die Qualität der fertigen Schweißverbindungen gemäß den Anforderungen der relevanten Vorschriften überprüft.]

114) Отжиг заключается в нагреве металла до определенной температуры, выдержке при этой температуре и затем медленном охлаждении вместе с печью.
[*Übersetzung*: Das Glühen besteht im Erhitzen des Metalls bis zu einer bestimmten Temperatur, dem Halten des Werkstücks unter dieser Temperatur für einen gewissen Zeitraum und seiner nachfolgenden langsamen Abkühlung gemeinsam mit dem Ofen.]

115) Сваркой соединяют однородные и разнородные металлы и их сплавы, металлы с некоторыми неметаллическими материалами (керамикой, графитом, стеклом и др.), а также пластмассы.
[*Übersetzung*: Durch Schweißen verbindet man gleich- und verschiedenartige Metalle und ihre Legierungen ebenso wie Metalle mit manchen nichtmetallischen Stoffen (Keramik, Graphit, Glas etc.) und Kunststoffe.]

Auch diese Beispiele weisen die bereits mehrfach als typisch bezeichneten linguistischen Merkmale wissenshaltiger Kontexte auf, z. B. die satzinitiale Position des Definiendums, Verwendung bestimmter grammatischer Formen des Verbs, eine hohe Anzahl von Nominalphrasen sowie das weitgehende Fehlen subjektiver oder relativierender lexikalischer Elemente. Hoch gerankte generische Sätze scheinen dagegen auch hier häufig eher lange Sätze mit tendenziell vielen Termini bzw. solche Sätze zu sein, die durch mindestens ein – möglichst zuverlässiges – Muster extrahiert wurden.
Insgesamt kann damit festgehalten werden, dass die in Kapitel 5 identifizierten linguistischen Merkmale wissenshaltiger Kontexte durchaus sinnvoll zu Extraktionszwecken verwendet werden können. In der Literatur genannte Wortarten wie Präpositionen und Relativpronomen sind in die hier beschriebenen Modellierungsaufgaben ebenso mit eingeflossen wie bestimmte grammatisch-strukturelle *features* (z. B. Termposition und syntaktische Realisierung, Passivgebrauch). Auch die Hypothese, dass KRCs zur Bildung von *clustern* neigen, scheint durch die positive Evaluation des Merkmals NEAR_HITS bestätigt. Im Zusammenspiel mit musterbezogenen Informationen können diese Merkmale wissenshaltige Kontexte zwar noch nicht so weit von generischen Sätzen unterscheiden, dass die Ausgabe eines entsprechenden Extraktionssystems in einem System für Endanwender direkt zum Einsatz kommen könnte, vielmehr war auf die Notwendigkeit einer inhaltlichen und sachlichen Überprüfung durch einen Terminologen bereits mehrfach verwiesen worden. Allerdings besteht auch Grund zu der Annahme, dass durch leistungsfähigere Filter- und Annotationsmechanismen bei der Korpusaufbereitung sowie eine elaboriertere statistische Modellierung der Daten noch deutliche Verbesserungspotentiale realisiert werden können.

6.3.2 Deutsch

6.3.2.1 Daten und Modellierung

In diesem Abschnitt soll das zuletzt erprobte Klassifikationsverfahren als diejenige Methode, welche die größten Verbesserungspotentiale verspricht, auch auf die deutschen Daten angewandt werden.

Dabei werden im wesentlichen dieselben Merkmale modelliert wie bereits für die russischen Daten, wobei das Merkmal GUT die *tags* NN, ADJA, VVPP und KON umfasst, das Merkmal SCHLECHT dagegen die *tags* ADV, KOUS, NE, PIAT, PPOSAT, PPER, PTKNEG, VMFIN und VVINF, die Zuordnung orientiert sich also an den Ergebnissen der im vorigen Kapitel durchgeführten linguistischen Analyse. Zusätzlich zu den für das Russische verfügbaren Informationen kann im Falle des Deutschen zudem auf zwei weitere Merkmalstypen zugegriffen werden, und zwar höherwertige *chunk*-Informationen, die von TreeTagger ausgegeben werden und anhand derer beispielsweise Satzanfänge beschrieben werden können, sowie Informationen zu den semantischen Klassen und Synsets von Prädikaten, die aus GermaNet abzurufen sind. Des Weiteren wird wie bereits im vorigen Kapitel das Merkmal PHOR modelliert. Tabelle 82 illustriert die Merkmalsannotation wiederum an einem Beispiel und zwar anhand des generischen Satzes „Parallel zu den Auftriebsläufern gibt es schon wesentlich länger die so genannten Widerstandsläufer.“

Merkmal	**Wert**
ID	4320
SENT_L	14
KRC	0
PATTERN_N	0
PATTERN_V	0
NEAR_HITS	1 (generischer Satz 4316 und KRC 4323)
NOUN_FIRST	0
VERB_FIRST	0
PREP_FIRST	1
ELSE	0
CHUNKS_N	5
GUT	3 (Auftriebsläufern (NN), genannten (VVPP), Widerstandsläufer (NN))
GUT_PROP	0,2
SCHLECHT	6 (Parallel (ADJD), es (PPER), schon (ADV), wesentlich (ADJD), länger (ADJD), so (ADV))
SCHLECHT_PROP	0,4
GUT_SCHLECHT_R	-0,2
BOOST_N	0
BOOST_PROP	0
STOP_N	1 (es)
STOP_PROP	0,1
BOOST_STOP_R	-0,1
PRED_CLASS_N	3 (geben: Allgemein, Besitz, Gesellschaft)
PRED_N	1
PRED_R	0,1
SYNSETS	5 (Beispielsätze: Es gibt heute weniger Bauern als vor 20 Jahren. Der Vater gibt seinem Sohn etwas Geld. Das Mädchen gab dem Jungen ihren Teddy. Sie gibt sich cool. Das Ensemble gab eine Vorstellung in der Stadt.)
BOOSTM_N	1 (geben)
BOOSTM_PROP	0,1
STOPM_N	0
STOPM_PROP	0
BOOSTM_STOPM_R	0,1
PHOR	0

ABSTAND	0
LAST	1
TERM_N	0
TERM_S	0
TERM_PO	0
TERM_DO	0
TERM_GA	0
TERM_PA	0
TERM_INI	0
PHRASE	0

Tabelle 82: Merkmalsannotation für einen deutschen Beispielsatz

Die Merkmale NOUN_FIRST (=erster *chunk* ist eine Nominalphrase), VERB_FIRST (=erster *chunk* ist eine Verbalphrase), PREP_FIRST (=erster *chunk* ist eine Präpositionalphrase) und ELSE (=keine der vorigen Varianten) kodieren Informationen zur Realisierung des Satzanfangs von KRC-Kandidaten in einer Serie einzelner Binärvariablen. Das Merkmal CHUNKS_N gibt dagegen an, wieviele *chunks* laut TreeTagger ein Satz enthält. Das Merkmal GUT_SCHLECHT_R ist die Differenz aus den Anteilswerten für die *features* GUT und SCHLECHT, wohingegen das Merkmal PRED_R angibt, wieviele Prädikate ein Satz bezogen auf seine Satzlänge enthält. Die Merkmale PRED_CLASS_N und SYNSETS kodieren die genannten Informationen aus GermaNet. Die Trainingsdaten für die Berechnung des Regressionsmodells umfassen die 339 KRCs, die im Trainingskorpus identifiziert wurden, sowie 339 zufällig ausgewählte generische Sätze.
Das im Ergebnis des Modellierungsprozesses errechnete AIC-Modell erreicht eine Devianz von 314,75 im Vergleich zu einer Nulldevianz von 939,91. Keiner der durchgeführten Tests deutet auf eine mangelhafte Modellanpassung oder eine Verletzung der Modellannahmen hin. Analog zu Tabelle 80 fasst Tabelle 83 die Modellparameter zusammen.

Merkmal	**Koeffizient**	**Standardfehler**	**Konfidenzintervall**	**Signifikanz**
PATTERN_V	2,7111	0,5396	[1,8529, 4,0523]	ja
NEAR_HITS1	1,0176	0,2951	[0,4437 1,6041]	ja
NOUN_FIRST1*	0,5709	0,3106	[-0,0298, 1,1918]	nein
GUT_SCHLECHT_R	2,9983	0,7846	[1,5014, 4,5879]	ja
BOOST_STOP_R	10,1615	3,0084	[4,4772, 16,3147]	ja
PRED_CLASS_N	0,1994	0,0837	[0,0353, 0,3653]	ja
TERM_N*	0,3418	0,1949	[-0,0392 0,7292]	ja
TERM_S1	0,8366	0,4044	[0,0423, 1,6335]	ja
TERM_PO1	1,2427	0,4469	[0,3784, 2,1388]	ja
TERM_GA1*	0,6734	0,4207	[-0,1510, 1,5041]	nein
TERM_INI1*	0,7368	0,3983	[-0,0430, 1,5247]	nein
LAST1	2,5756	0,8587	[1,0802, 4,5720]	ja

Tabelle 83: Parameter des deutschen AIC-Modells

Hier wurden ausschließlich positive Merkmale, also solche *features* selektiert, die auf das Vorliegen eines KRCs hindeuten. Gemäß dem Signifikanztest auf Grundlage der Devianzdifferenzen sind zudem alle Merkmale bis auf NOUN_FIRST, TERM_GA und TERM_INI statistisch signifikant, wobei allerdings das Konfidenzintervall für das Merkmal TERM_N, das den Wert 0 beinhaltet, Zweifel an der Signifikanz auch dieses Merkmals aufkommen lässt. Die entsprechenden Merkmale sind in der Tabelle durch * markiert.

6.3.2.2 Ergebnisse

Das im vorigen Abschnitt beschriebene deutsche Regressionsmodel wurde für die Analyse der 10.696 Sätze des deutschen Testkorpus‘ verwendet, unter denen 287 extrahierbare KRCs identifiziert wurden. Im Vergleich zu den russischen Daten, die in den vorigen Abschnitten analysiert wurden, ist damit ein deutlich geringerer Anteil von KRCs im Testkorpus zu identifizieren, wodurch die Klassifikationsaufgabe erschwert wird. Abbildung 35 stellt die auf den einzelnen Abdeckungsniveaus erreichten Ergebnisse, die mit Hilfe der logistischen Modellierung erzielt werden, wenn absteigend nach den errechneten Regressionskoeffizienten sortiert und ein entsprechender Schwellenwert für eine Klassifikationsentscheidung angesetzt wird, den Ergebnissen der musterbasierten Extraktion sowie einem *baseline*-Ranking nach der Anzahl der Termini pro Satz gegenüber, wobei für Letzteres wiederum keine globalen, sondern gruppenbezogene Präzisionswerte angegeben werden. Auch Kurven für das jeweilige f-Maß, das bei musterbasierter bzw. logistischer Extraktion erreicht wird, sind angegeben.

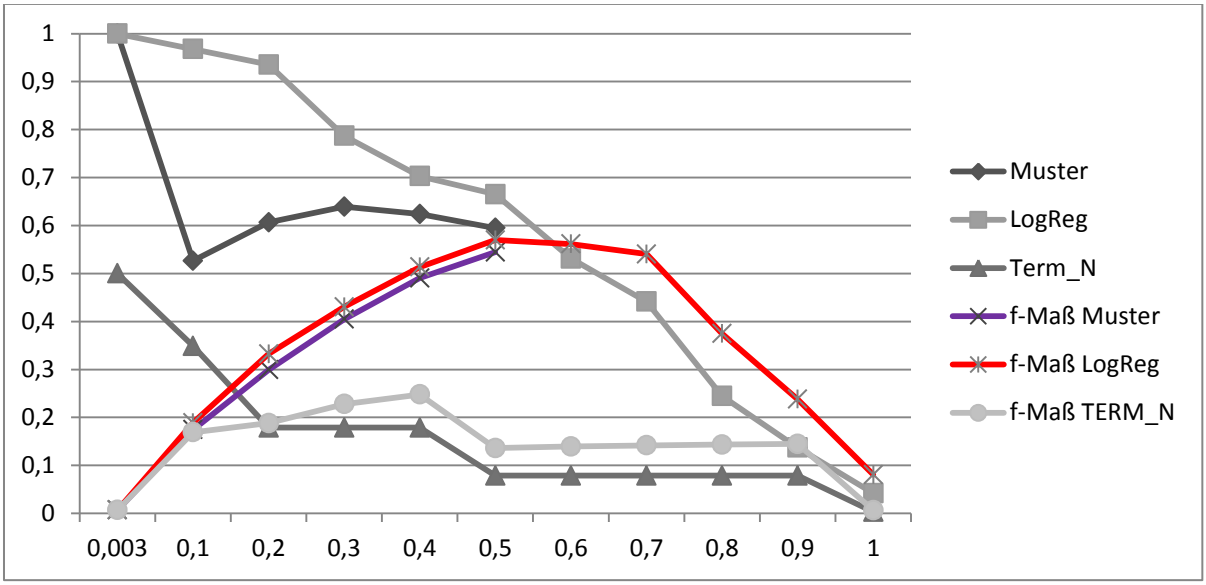


Abbildung 35: Präzisions-Abdeckungs-Kurve für die Extraktion deutscher KRC-Kandidaten des Sub-TKD mit dem erweiterten Regressionsmodell im Vergleich zu zwei Baselines

Die Abbildung zeigt, dass mit der Regressionsmethode im Vergleich zur musterbasierten Extraktion, mit der für das Deutsche – im Gegensatz zu den russischen Daten – eine relativ hohe Performanz erreicht wurde, insbesondere im oberen Bereich des Rankings, d. h. für niedrigere Abdeckungsniveaus, eine deutlich höhere Präzision erreicht wird. Bei einer Abdeckung von etwa 0,55 sinkt das Präzisionsniveau allerdings auf das der musterbasierten Extraktion ab, wobei der Vorteil der Regressionsmethode jedoch darin besteht, dass eine Bewertung *aller* Korpussätze durchgeführt wird. In Verbindung mit der zumindest für die obere Sektion der gerankten KRC-Kandidaten deutlich höheren Präzision kann damit auch für solche Abdeckungsbereiche, die durch die musterbasierte Extraktion nicht mehr erfasst werden, noch eine annehmbare Präzision und insgesamt ein höheres f-Maß von 0,57 (auf einem Abdeckungsniveau von 0,5 im Vergleich zu einem f-Maß von 0,56 bei 60-prozentiger Abdeckung) erreicht werden. Erstere Angabe entspricht einer Klassifikationsgenauigkeit von 98%. Auch auf diesem Abdeckungsniveau werden wiederum KRCs extrahiert, welche mit Hilfe der musterbasierten Methode allein nicht identifiziert werden können (sic!):

116) Eine Sonderrolle nehmen die Magischen Augen ein , deren primärer Zweck nicht die Verstärkung oder Manipulation von Signalen darstellt , sondern die Umsetzung einer Signalgröße in ein entsprechendes optisches Äquivalent .

117) Decoys nehmen das aktive Sonarsignal des angreifenden Torpedos auf und senden es als Scheinecho zurück , um die Waffe anzulocken und vom richtigen Ziel abzulenken .

118) Zentralschmiersysteme sorgen dafür , dass der Schmierstoff in der entsprechenden Menge und in bedarfsgerechten Intervallen den zu schmieren - den Stellen einer Windkraftanlage zugeführt wird (Bild 1) .

Auch diese KRCs scheinen wiederum diejenigen Merkmale zu illustrieren – z. B. Terminus in Subjektposition, Verzicht auf Schachtelsätze, Verwendung von Relativsätzen, hoher Anteil an Nominalphrasen –, die bereits in den vorigen Abschnitten als KRC-typisch identifiziert wurden. Eine Interpretation der Regressionskoeffizienten des errechneten Modells im Hinblick auf die Relevanz konkreter Merkmale ist problematisch, da insbesondere die Möglichkeit direkter Vergleiche beispielsweise mit dem im vorigen Kapitel beschriebenen Modell nicht gegeben ist. Tendenziell kann indes festgestellt werden, dass durch die Hinzunahme zusätzlicher Merkmale die Bedeutung der wortartenbezogenen Merkmale – gemessen an ihren Koeffizienten – zurückzugehen scheint, wohingegen beispielsweise die Koeffizienten der lexikalischen und terminologischen Merkmale im Vergleich zu dem in Kapitel 5 berechneten Modell gleich zu bleiben scheinen bzw. für die lexikalischen Merkmale sogar noch gesteigert werden. Wird wiederum ein Testdatensatz von 100 Sätzen aus dem Sub-TKR, der 11 KRCs und 89 generische Sätze enthält, herangezogen, um die Auswirkungen systematischer Merkmalsmanipulationen zu untersuchen, zeigt sich das in Tabelle 84 zusammengefasste Bild.

Merkmal	**Merkmal**	**Mittelwert Vorhersage**	**Signifikanz**
PATTERN_V			
setting0	0,1500	0,2287	signifikant für setting0 vs. setting2, setting1 vs. setting2
setting1	0,0900	0,2270	
setting2	1,1500	0,6244	
NEAR_HITS			
setting0	0: 60, 1: 40	0,2287	signifikant für setting0 vs. setting2, setting1 vs. setting2
setting1	0: 100, 1: 0	0,1758	
setting2	0: 0, 1: 100	0,2866	
NOUN_FIRST			
setting0	0: 30, 1: 70	0,2287	signifikant für setting1 vs. setting2
setting1	0: 100, 1: 0	0,1895	
setting2	0: 0, 1: 100	0,2434	
GUT_SCHLECHT_R			
setting0	0,2410	0,2287	signifikant für setting1 vs. setting2
setting1	0,1462	0,2084	
setting2	0,3338	0,2625	
BOOST_STOP_R			
setting0	-0,0270	0,2287	signifikant für alle drei Vergleiche
setting1	-0,1213	0,1687	
setting2	0,0663	0,3361	

PRED_CLASS_N			
setting0	2,2700	0,2287	nicht signifikant
setting1	1,4900	0,2118	
setting2	3,2700	0,2498	
TERM_N			
setting0	0,2800	0,2287	nicht signifikant
setting1	0,0500	0,2181	
setting2	1,2800	0,2660	
TERM_S			
setting0	0: 90, 1: 10	0,2287	signifikant für setting0 vs. setting2, setting1 vs. setting2
setting1	0: 100, 1: 0	0,2198	
setting2	0: 0, 1: 100	0,3198	
TERM_PO			
setting0	0: 96, 1: 4	0,2287	signifikant für setting0 vs. setting2, setting1 vs. setting2
setting1	0: 100, 1: 0	0,2234	
setting2	0: 0, 1: 100	0,3830	
TERM_GA			
setting0	0: 94, 1: 6	0,2287	signifikant für setting0 vs. setting2, setting1 vs. setting2
setting1	0: 100, 1: 0	0,2207	
setting2	0: 0, 1: 100	0,2990	
TERM_INI			
setting0	0: 91, 1: 9	0,2287	signifikant für setting0 vs. setting2, setting1 vs. setting2
setting1	0: 100, 1: 0	0,2223	
setting2	0: 0, 1: 100	0,3095	
LAST			
setting0	0: 8, 1: 92	0,2215	signifikant für setting0 vs. setting1, setting1 vs. setting2
setting1	0: 100, 1: 0	0,0901	
setting2	0: 0, 1: 100	0,2553	

Tabelle 84: Auswirkungen der systematischen Variation von Merkmalswerten auf einer Stichprobe aus dem Sub-TKD

Zunächst zeigt sich, dass die Variation der Werte für die Merkmale PRED_CLASS_N und TERM_N keine signifikanten Veränderungen der Klassifikationsergebnisse nach sich zieht, was mit den Ergebnissen der Signifikanztests für das Modell übereinstimmt – die Informationen aus GermaNet und Angaben zur Anzahl der Termini im Satz scheinen für die Klassifikation von wissenshaltigen Kontexten mit dem hier verwendeten Modell mithin nur von untergeordneter Bedeutung zu sein. Die in dem Merkmal GUT_SCHLECHT_R zusammengefassten Eigenschaften zeitigen signifikante Veränderungen nur dann, wenn mindestens zwei *tokens* variiert werden (setting1 vs. setting2), was als eher drastische Variation gelten kann, wohingegen Boost- und Stopwörter einen größeren Effekt zu haben scheinen. Morphologische Merkmale scheinen im Vergleich zum Russischen von untergeordneter Bedeutung zu sein, was daran liegen könnte, dass für das Deutsche das als wesentlich angesehene Merkmal „Passiv“ auf morphologischer Ebene nicht modelliert werden kann – für die Untermauerung einer solchen Aussage wären allerdings eingehendere Studien erforderlich. Ebenso wie im Fall des Russischen schlagen sich Variationen des Merkmals LAST stets in hochsignifikanten Ergebnissen nieder.

6.3.2.3 Zwischenfazit

Auch für das Deutsche kann durch ein einfaches Regressionsverfahren, das neben korpusbasierten Merkmalen wissenshaltiger Kontexte, wie sie im letzten Kapitel erarbeitet wurden, auch musterbezogene Informationen in die Modellierung einbezieht, eine Verbesserung der Extraktionsperformanz erreicht werden. Dass diese Steigerung der Extraktionsqualität im Vergleich zum Russischen geringer ausfällt, kann dabei angesichts der für das Deutsche deutlich besseren Ausgangsperformanz insofern als erwartbares Ergebnis gelten, als auf Grundlage eines geringeren Ausgangsniveaus Steigerungen bereits durch geringfügige Verbesserungen realisiert werden können. Allerdings ist auch hier davon auszugehen, dass durch Anwendung elaborierterer Modellierungs- und Lernmethoden noch über die hier referierten hinausgehende Verbesserungen erzielt werden können.
Generell scheinen die für das Deutsche erzielten Ergebnisse die Hypothesen zum Nutzen linguistischer Eigenschaften für die Identifikation wissenshaltiger Kontexte für das Gros der in Kapitel 5 erarbeiteten Merkmale zu bestätigen, wobei in sprachübergreifender Sicht wiederum Parallelen zum Russischen deutlich werden. Beispielsweise favorisieren auch deutsche KRCs Termini in Subjektposition, zudem scheinen KRCs auch in dem hier untersuchten Datensatz in Gruppen aufzutreten, was die Selektion des Merkmals NEAR_HITS zeigt. Darüber hinaus erweisen sich morphologische und lexikalische Merkmale als kennzeichnend für KRCs, wobei indes morphologische Merkmale im Vergleich zum Russischen von tendenziell geringerer Bedeutung zu sein scheinen. Musterbezogene Informationen insbesondere zur Zuverlässigkeit von Extraktionsmustern erscheinen damit zwar als wichtiges, jedoch bei Weitem nicht als einziges Kriterium, anhand dessen wissenshaltige Kontexte von generischen Aussagen unterschieden werden können. Andere Merkmalsgruppen scheinen dagegen keine oder nur eine geringe Bedeutung für die Klassifikation von KRCs zu haben.

6.4 Qualität automatisch extrahierter wissenshaltiger Kontexte aus Web-Korpora und Perspektiven der Anreicherung von Terminologiedatenbanken mit korpusbasierter Information

In den letzten Abschnitten sind Verfahren für die automatische Extraktion wissenshaltiger Kontexte aus *web*-Korpora entwickelt, beschrieben und evaluiert worden. In Bezug auf die Extraktionsperformanz war dabei darauf hingewiesen worden, dass bislang eine direkte Integration von Extraktionsergebnissen in terminologische *workflows* nicht möglich scheint, da in den letzten Abschnitten zwar ermutigende, allerdings für eine ungeprüfte Übernahme in terminologische Ressourcen noch keine ausreichende Güte aufweisende Ergebnisse erzielt wurden. Im Zusammenhang mit der Erstellung des Goldstandards war zudem festgestellt worden, dass sich eine terminologische Korrektheitsprüfung in jedem Fall *an den Extraktionsschritt anschließen* muss, zumal im Zuge der Identifikation falscher Positive nur solche KRC-Kandidaten aus dem Goldstandard entfernt worden waren, die ganz offensichtliche inhaltliche Mängel oder Widersprüche aufwiesen. In diesem Abschnitt sollen diese Aspekte im Hinblick auf die einleitend aufgeworfene Frage, inwieweit Extraktionsergebnisse tatsächlich Nutzungsperspektiven für die praktische Terminologiearbeit bieten, etwas ausführlicher diskutiert werden. Aus dieser Diskussion sollen schließlich auch Desiderata für zukünftige Forschungsarbeiten abgeleitet werden.

Die zu berücksichtigenden Gesichtspunkte werden zunächst anhand von Extraktionsergebnissen illustriert. Die Sätze 119)-133) wurden mit Hilfe der in Abschnitt 6.2.3.1.3 beschrieben Methode aus dem Sub-EKD extrahiert. Überflüssiges sprachliches Material in den Sätzen aufgrund einer unvollkommenen Satzsegmentierung wurde aus darstellungspraktischen Gründen entfernt, eine darüber hinaus gehende Normalisierung der sprachlichen Gestaltung wurde jedoch nicht vorgenommen.

119) Die Anlagen besitzen auch eine Notstromversorgung , um bei Netzausfall ein sicheres Abschalten (Blätter in Segelstellung drehen und / oder bremsen) zu gewährleisten .

120) Eine Sturmregelung in neueren Anlagen erlaubt den reduzierten sicheren Betrieb der Anlage bei fast jeder Windgeschwindigkeit .

121) Drehzahlvariable pitchgeregelte Anlagen Hier wird zwischen zwei Betriebszuständen unterschieden : der Drehzahlregelung im Teillastbetrieb (Momentenregelung) und der Drehzahlregelung im Volllastbetrieb (Pitchregelung) .

122) Hinsichtlich der Bauformen unterscheidet man zwischen Anlagen mit horizontaler und vertikaler Achse .

123) Eine Windenergieanlage (WEA) wandelt Windenergie in elektrische Energie um und speist diese zumeist in das öffentliche Stromnetz ein .

124) Windenergieanlagen können in allen Klimazonen , auf See und in allen Landformen (Küste , Binnenland , Gebirge) zur Gewinnung elektrischen Stroms eingesetzt werden .

125) Moderne Windenergieanlagen zur Stromerzeugung sind aerodynamisch angetriebene Anlagen .

126) Die gesamte zur Energieumwandlung dienende Funktionseinheit wird als Windenergieanlage oder-konverter bezeichnet .

127) Bei der Windenergienutzung wird zwischen Windgeneratoren mit horizontaler (Horizontalachsenkonverter) und vertikaler (Vertikalachsenkonverter) Drehachse unterschieden , wobei letztere bisher keine Rolle im kommerziellen Sektor spielen .

128) Die Energiegewinnung durch Windgeneratoren stellt neben Photovoltaik , Biogas und Wasserkraft eine der sichersten und umweltfreundlichsten Methoden dar .

129) Eine Windkraftanlage , auch Windkonverter genannt , besteht im Wesentlichen aus dem Windrad oder Rotor und dem Generator , der mit dem Rotor über eine gemeinsame Achse verbunden ist .

130) Windkraftanlagen mit vertikaler Rotationsachse gibt es (unter anderem) in der Ausführung als Savonius-Rotor oder Darrieus-Rotor .

131) Es gibt im Wesentlichen zwei Bauformen von Windkraftanlagen : solche mit senkrechter Drehachse und solche mit waagrechter Achse .

132) Es wird bei Windkraftanlagen zwischen offshore und onshore Anlagen unterschieden .

133) Von der Stellung der Windräder kann man solche unterscheiden , die entweder in Luv (der Wind bläst von vorne auf die Gondel , bzw . die Rotorblätter) oder in Lee (der Wind bläst von hinten auf die Gondel bzw . die Rotorblätter) stehen .

Jedes dieser Beispiele vermittelt spezifische Informationen zu Windkraftanlagen, und zwar zu Komponenten bzw. Elementen von Windkraftanlagen, Betriebszuständen, Einsatzmöglichkeiten und grundlegenden Konstruktionsprinzipien, Bauformen von Windkraftanlagen, ihrer Funktion sowie möglichen Standorten. Keiner der Beispielsätze liefert eine vollumfängliche Beschreibung von Windkraftanlagen oder eine intensionale Definition. Als problematischer als diese beiden Aspekte erscheinen jedoch die folgenden Gesichtspunkte:

- *Verwendung von Synonymen*: In den Beispiel-KRCs werden zahlreiche Synonyme bzw. alternative Benennungen für den interessierenden Terminus verwendet, z. B. Anlage, WEA, Windenergieanlage, Windenergiekonverter, Windgenerator und Windrad. Sollen tatsächlich umfangreiche terminologische Informationen aus Korpora extrahiert werden, ergibt sich aus dieser Inkonsistenz auf der Ebene der Benennungen die Notwendigkeit, Synonyme einander zuzuordnen und Vorzugsbenennungen festzulegen. Sofern hierfür nicht auf umfangreiche Datenbankbestände zugegriffen werden kann, setzt dies auf technischer Seite die korpusbasierte Extraktion und das *mapping* von Termini und Benennungsvarianten voraus.
- *Redundanz der Information*: Redundanz nicht im Bierwischschen (Kapitel 2) Sinn, sondern im Sinne inhaltlicher Wiederholungen kann auf mehreren Ebenen auftreten, und zwar einerseits auf der Ebene der Definienda selbst, d. h. zu hochfrequenten, tendenziell allgemeineren Termini wird in einem *web*-Korpus – zumindest dann, wenn es so erstellt wurde, wie die hier verwendeten Korpora – in der Regel mehr Information zu finden sein als zu hochspezialisierten Termini. Andererseits scheinen auch bestimmte Informationstypen häufiger in KRCs genannt zu werden, in vorstehenden Beispielen etwa die Information zu unterschiedlichen Bauformen von Windkraftanlagen. Aus extraktionspraktischer Sicht ist dies insofern problematisch, als wiederholte bzw. sehr ähnliche Inhalte gefiltert und darüber hinaus größere Datenmengen untersucht werden müssen, damit auch stärker spezialisierte Informationen identifiziert werden können. Weitere Desiderata sind mithin die Erstellung umfangreicherer Datensätze, womit sowohl Korpora als auch entsprechend annotierte Goldstandard-Ressourcen gemeint sind, sowie die Nutzung bzw. Entwicklung verbesserter Filtertechniken.
- *Fehlende oder unbekannte Systematik der begrifflichen Beschreibung*: Es war bereits darauf hingewiesen worden, dass Aspekte der Wissensvermittlung in Texten ein kontextuelles Phänomen darstellen und die Beschränkung der hier durchgeführten Untersuchung auf die Ebene des Einzelsatzes eine im Wesentlichen sachlichen Zwängen geschuldete Einschränkung darstellt. Tatsächlich machen die oben angeführten Beispiele deutlich, dass die Mehrzahl der aus Korpora extrahierten KRCs nur Informationsbruchstücke liefert, die im Textzusammenhang erst in Kombination mit anderen Sätzen ein vollständigeres Bild geben. Inwieweit aus unterschiedlichen Korpustexten extrahierte Bruchstücke nun miteinander zu

vollumfänglichen Begriffsbeschreibungen kombiniert werden können (oder ob bestimmte Ausschlusskriterien bzw. sachliche *constraints* existieren), ist momentan nur durch zusätzliche Terminologiearbeit festzustellen. Das Gros der extrahierbaren KRCs liefert also tatsächlich – neben benachbarten Begriffen und Benennungsvarianten – bestenfalls Ansatzpunkte für die Formulierung von Definitionen und Begriffsbeschreibungen. Aus dieser Feststellung ergibt sich ein theoretisches Desideratum, nämlich dasjenige der Entwicklung von Modellen und Werkzeugen für die Beschreibung von Phänomenen der sprachgebundenen Wissensvermittlung auf diskursiver Ebene. Entsprechende Ansätze – dies hat etwa die referierte Diskussion in Roelcke (2013, Kapitel 3) deutlich gemacht – scheinen bislang nicht zu existieren.

- *Abhängigkeit der KRC-Qualität von der Textqualität*: Auch der Zusammenhang zwischen den Eigenschaften von Texten, aus denen KRCs extrahiert werden, und der Qualität der Extraktionsergebnisse selbst sowie der Zusammenhang zwischen Textsorte und KRC-Frequenz waren bereits in Ansätzen diskutiert worden. Aus *web*-Texten bestehende Korpora waren dabei als Datengrundlage für die hier beschriebenen Experimente gewählt worden, weil Übersetzer und Dolmetscher bei der Terminologierecherche nicht immer Zugriff auf höherwertige Ressourcen, z. B. Fachbücher und Fachwörterbücher in elektronischer oder gedruckter Form, haben. Neben den bereits genannten, aus einer qualitätsbezogenen Sicht kritischen Aspekten ist zudem in Bezug auf Online-Quellen – auch hierauf wurde hingewiesen – mit einem höheren Anteil sprachlicher und inhaltlicher Schwächen zu rechnen. Ein grundsätzliches methodologisches Problem scheint darüber hinaus auch darin zu bestehen, dass der für die Korpuserstellung eingesetzte *crawler* „Babouk“ – die Analysen in Kapitel 4 dürften dies gezeigt haben – neben brauchbaren Texten auch große Mengen irrelevanten Materials selektiert hat. Insbesondere war zumindest zum Zeitpunkt der Korpuserstellung keine Auswahl bestimmter Texttypen oder Textsorten mit Hilfe von „Babouk“ möglich. Die durchgeführten Experimente und Analysen geben jedoch Grund zu der Annahme, dass durch eine gezielte Textauswahl die Qualität der extrahierten Information in Bezug auf die hier genannten Punkte gesteigert werden kann, womit schließlich ein weiteres eher technisches Desideratum formuliert werden kann, nämlich die Entwicklung von Methoden für die automatische oder zumindest semi-automatische Erstellung stärker – nämlich in Bezug auf die gesammelten Texttypen und -sorten – spezialisierter Korpora, wobei die Automatisierbarkeit des Prozesses angesichts der Vielfalt der von Übersetzern und Dolmetschern zu bedienenden Sprachenpaare und Fachgebiete eine wesentliche Anforderung darstellt.

Über diese beispielorientierte Diskussion hinaus ist in Bezug auf Extraktionsergebnisse auch eine Qualitätsbewertung auf Grundlage der in Kapitel 2 vorgestellten Typologie wissenshaltiger Kontexte nach Bierwisch und Kiefer (1969) möglich. Auch die grundlegenden Ergebnisse dieser Bewertung sollen anhand von Beispielen diskutiert werden (sic!):

134) Der Prozess bei dem sich abgestorbenes organisches Material zuerst in Torf dann in Braunkohle und zuletzt in Steinkohle umwandelt wird als Inkohlung bezeichnet und ist durch Zunahme des relativen Kohlenstoffanteils gekennzeichnet .

Dieses Beispiel repräsentiert einen KRC, der aus dem Sub-TKD extrahiert wurde, die Textquelle ist Wikipedia[138]. Gemäß den in Kapitel 2 referierten linguistischen Klassifikationen handelt es sich um einen definitorischen Kontext. Durch das in den vorigen Abschnitten beschriebene Rankingverfahren wurden diesem Satz ein sehr hoher Ranking-Wert (0,9999) sowie die sechzehnte Ranking-Position zugeordnet. Wird dagegen das von Bierwisch und Kiefer (1969) genannte Kriterium der Vollständigkeit auf alle in Kapitel 2 definierten semantischen Zielrelationen bezogen (nicht alle dort genannten Relationen sind auf den Begriff „Inkohlung“ anwendbar, die Relation Kausalität beispielsweise aber sehr wohl), kann von einer Vollständigkeit der durch den KRC vermittelten Information keine Rede sein. Tatsächlich kann vielmehr davon ausgegangen werden, dass in einem Übersetzungskontext auch Aspekte des Begriffs „Inkohlung“ recherchiert werden müssen, die durch den KRC nicht beschrieben werden (z. B. über welchen Zeitraum sich der Prozess der Inkohlung erstreckt oder durch welche Faktoren er ausgelöst oder begünstigt wird). In diesem Sinne handelt es sich bei allen extrahierten KRCs um „partielle“ Kontexte nach Bierwisch und Kiefer. Ein deutlich schlechter als Beispiel 134) gerankter KRC (Ranking-Wert 0,8040, Position 322) ist dagegen Beispiel 116). Dieser KRC vermittelt im Gegensatz zu 134) eher als „redundant“ zu charakterisierende Informationen (und zwar eine recht unspezifische Beschreibung des Verwendungszwecks). In diesem Sinne „redundant“ ist indes ein großer Teil der aus dem Sub-EKD extrahierten KRCs, so findet sich unter den Beispielen 119)-133) lediglich ein KRC, der nicht „redundant“ ist (Beispiel 123)), sondern Informationen liefert, die sich auch in einer aristotelischen Definition finden könnten (freilich nennt der Kontext kein *genus proximum*).

Insgesamt bestätigt die qualitätsbezogene Analyse damit wiederum die Hypothese, dass automatisch extrahiertes Korpusmaterial bestenfalls Ansatzpunkte für die terminologische Begriffsbeschreibung liefert. Deutlich wird aber auch, dass eine „vollständige“ Begriffsbeschreibung im hier intendierten Sinne auch durch klassische – intensionale oder extensionale – Definitionen nicht erreicht werden kann, so dass der eingangs zitierten Kritik Gruntar Jermols (2010) Recht zu geben ist. Vor diesem Hintergrund bleibt die durch wissenshaltige Kontexte bzw. deren Kombination repräsentierte Möglichkeit der Erstellung umfangreicherer Begriffsbeschreibungen nach wie vor relevant, wenngleich die automatische Extraktion und Repräsentation semantischer Relationen zwischen Begriffen zumindest aus terminologischer Sicht momentan nicht praktikabel scheint, die Inklusion entsprechender Informationen in terminologische Ressourcen vielmehr nach wie vor manuelle Arbeitsschritte voraussetzt. Aus der angestrebten Tiefe der terminologischen Beschreibung ergibt sich zudem das Desideratum der Entwicklung von Modellen und Methoden für die strukturierte und an unterschiedliche Informationsbedürfnisse anpassbare Präsentation begriffsbezogener Information, d. h. eine Weiterentwicklung der bislang in der Regel statischen Datenkategorien für die Angabe kontextueller und definitorischer Information, denn nur durch strukturierte und flexible Formen der Repräsentation lässt sich die Fülle der aus Korpora extrahierbaren Information beherrsch- und darstellbar machen.

Im Ergebnis lässt die Diskussion dieser Qualitätsaspekte bzw. vielmehr -probleme die Nutzung automatisch extrahierter KRCs für die praktische Terminologiearbeit zwar nicht als aussichtslos erscheinen, macht aber deutlich, dass eine vollständige Automatisierung der Erstellung von Begriffsbeschreibungen noch in weiter Ferne liegt und an zahlreiche Voraussetzungen gebunden ist. Gegen eine Beschränkung der Terminologiearbeit auf den Umgang mit manuell erstellten oder automatisch aus frei zugänglichen Ressourcen extrahierten klassischen Definitionen spricht jedoch – dies sei hier noch einmal betont – insbesondere der Umstand, dass die Inklusion wissenshaltiger

[138] http://de.wikipedia.org/wiki/Energiequelle, abgerufen am 03.04.2014.

Kontexte in terminologische Begriffsbeschreibungen die Möglichkeit einer vollständigeren Dokumentation zwischenbegrifflicher semantischer Relationen und damit neben den in der Einleitung genannten Aspekten auch eine Erhöhung der Nutzungsqualität terminologischer Ressourcen aus der Sicht von Übersetzern und Dolmetschern und damit einer relevanten Nutzergruppe verspricht. Forschungsarbeiten in diesem Bereich sollten also weitergeführt werden.

6.5 Fazit

In diesem Kapitel wurden musterbasierte und hybride Ansätze zur Extraktion wissenshaltiger Kontexte aus deutschen und russischen *web*-Korpora entwickelt, beschrieben und evaluiert. Damit wurde der einleitend geforderten multilingualen Perspektive Rechnung getragen. Darüber hinaus wurden Extraktionsexperimente auf Testkorpora durchgeführt, die in Bezug auf ihre inhaltliche Zusammensetzung nicht mit den Entwicklungskorpora übereinstimmten und insbesondere andere Fachgebiete abdeckten, so dass die Generalisierbarkeit von Extraktionsmethoden zumindest in Ansätzen – die hier ausgewerteten Korpora sind immer noch zu klein – überprüft werden konnte.

Für das Deutsche wurden insgesamt 82, für das Russische dagegen nur 54 Extraktionsmuster implementiert und evaluiert, wobei auf den jeweils zwei Korpora für beide Sprachen f-Maße zwischen 0,4 und 0,6 erreicht wurden. Im Vergleich zu den Ergebnissen der in Kapitel 3 zitierten Studien erscheinen diese Resultate als kompetitiv, im Hinblick auf direkte praktische Anwendungen jedoch als zu niedrig, zumal die im vorigen Abschnitt geleistete Diskussion von Qualitätsaspekten gezeigt haben dürfte, dass praktische Anwendungen der KRC-Extraktion noch an eine Reihe weiterer, nicht ohne Weiteres realisierbare Voraussetzungen gebunden sind. Nichtsdestotrotz kann die Fortführung der Forschungsarbeit in diesem Bereich weiterhin als vielversprechend gelten.

In Bezug auf die Performanz musterbasierter Extraktionsmethoden und insbesondere im Hinblick auf die Frage nach der Übertragbarkeit von Extraktionsmustern auf ungesehenes, inhaltlich abweichendes Material konnten zwar für das Deutsche stabile Werte erreicht werden, für das Russische wurde jedoch ein drastischer Leistungsabfall beim Übergang vom Entwicklungs- auf das Testkorpus beobachtet, und zwar sowohl auf der Ebene der Abdeckung als auch auf jener der Präzision. Als Ursachen hierfür wurden neben den auch für das Deutsche festzustellenden Fehlerursachen fehlende Muster, zu eng formulierte Muster und Annotationsfehler auch die große Vielfalt russischer Sprachstrukturen einschließlich einiger spezifischer Konstruktionstypen (z. B. Schwierigkeit der Extraktion verbloser Bindestrich-Sätze) genannt. Zwar konnten für das Russische durch eine verbesserte Term-Erkennung sowie eine theoriebasierte Erweiterung der Suchmustermenge zumindest auf dem Testkorpus deutliche Verbesserungen erzielt werden. Allerdings zeigte eine Analyse der Performanz individueller Suchmuster – und hier bestätigten die russischen Ergebnisse wiederum die für das Deutsche erzielten –, dass nur wenige der manuell formulierten Extraktionsmuster auf beiden Korpora stabile Ergebnisse erbrachten, eine Beschränkung der Extraktion auf diese Muster aber aus Gründen der Abdeckung nicht ratsam sei. Die Übertragbarkeit manuell spezifizierter Suchmuster kann mithin zumindest für das hier untersuchte Datenmaterial in Zweifel gezogen werden. Ableiten lässt sich aus dem Ergebnis aber, dass linguistische Extraktionsmuster möglichst in einer formalisier- und kontrollierbaren Art und Weise aus größeren Datenmengen erlernt werden sollten – beispielsweise in Form lexikalisierter Dependenzstrukturen. Ob auf diese Weise Musterkataloge generiert werden können, die den von Hearst (1992, Abschnitt 3.4.1) genannten Qualitätskriterien in stärkerem Maße entsprechen als die hier genutzten Muster, ist allerdings noch zu untersuchen. Vorerst kann festgehalten werden, dass

linguistische Extraktionsmuster zwar Indizien für das Vorliegen eines KRCs, aber keine zuverlässigen Extraktionskriterien darstellen.

Im Anschluss an die Erprobung musterbasierter Extraktionsverfahren wurden wiederum Experimente mit logistischer Regression durchgeführt. Ziel dieser Experimente war es dabei insbesondere zu eruieren, inwieweit die im letzten Kapitel identifizierten quantitativen Merkmale wissenshaltiger Kontexte zu Extraktionszwecken genutzt werden können, da auf diese Weise nicht nur die Relevanz der erhobenen Merkmale überprüft, sondern auch Möglichkeiten der Integration von symbolischen und ML-Verfahren diskutiert werden können. Am vielversprechendsten schien dabei das Verfahren der Integration musterbezogener Informationen in die Klassifikation, mit dem für das beste *setting* auf dem russischen Testkorpus eine Performanzsteigerung von mehr als 10% (bezogen auf das beste erreichte f-Maß) erzielt werden konnte. Hiermit wurde ein Verfahren entwickelt, das nach Kenntnis der Autorin in der Literatur bislang noch nicht beschrieben wurde, einige Schwächen anderer Kombinationstypen jedoch überwindet. Für das Deutsche fiel die erzielte Leistungssteigerung deutlich geringer aus, wobei jedoch das höhere Ausgangsniveau zu berücksichtigen ist. In Bezug auf die Frage, ob die quantitativ erhobenen linguistischen Merkmale wissenshaltiger Kontexte die Extraktion von KRCs erleichtern, kann mithin festgestellt werden, dass dies der Fall ist, bei den erhobenen Merkmalen scheint es sich also tatsächlich um relevante *features* zu handeln. Gleichzeitig zeigen sich auch hier wieder dahingehend interlinguale Parallelen, dass terminologische (syntaktisch-positionelle), grammatische (bzw. morphologische) und lexikalisch-strukturelle Gesichtspunkte – bei scheinbar leichten interlingualen Unterschieden in der Gewichtung dieser Merkmale – für die logistische Modellierung von Bedeutung sind. Für das Russische hat sogar bereits die einfache logistische Modellierung in Abschnitt 6.3.1.1 ohne jede Anwendung von Suchmustern für das Sub-TKR eine Performanz im Bereich der musterbasierten Extraktionsgüte erbracht. Insgesamt bestätigen die hier präsentierten Resultate nicht nur den bereits zitierten und auch von Cramer (2011) replizierten Befund starker Schwankungen in der Performanz individueller Extraktionsmuster, sondern auch ihre Aussage (ebd.: 100), dass Kombinationen symbolischer mit statistischen Methoden in der Regel zu besseren Ergebnissen führen, wobei allerdings in der Regel nach wie vor eher mittelmäßige Ergebnisse erzielt werden. Festgehalten werden kann damit schließlich auch, dass die hier vorgestellte Kombination recht unterschiedlicher Suchmuster mit quantitativ erhobenen KRC-Merkmalen dem in dieser Arbeit verfolgten breiten Ansatz – die zitierten Beispiele extrahierter KRCs machen dies deutlich – der Beschreibung wissenshaltiger Kontexte eher gerecht zu werden scheinen als ein Ansatz, der sich von vornherein nur auf wenige sprachliche Strukturen (z. B. *Nomen-Kopula*-Nomen) konzentriert. Andererseits macht die eingenommene quantitative Perspektive auch deutlich, dass ein großer Anteil der zumindest in den hier untersuchten Daten aufgefundenen KRCs tatsächlich durch recht einfache Merkmalskonfigurationen beschrieben werden kann.

Die diskutierten qualitativen Aspekte schließlich verweisen darauf, dass zusätzlich zu den bereits mehrfach angesprochenen psycholinguistischen sowie den nunmehr hinlänglich diskutierten Performanzgesichtspunkten auch neue Fragestellungen in den Gesichtskreis der KRC-bezogenen Forschung treten müssen, so etwa Aspekte der Repräsentation vielfältiger begrifflicher Informationen in terminologischen Ressourcen.

7. Zusammenfassung und Ausblick

7.1 Ergebnisse der Arbeit

In dieser Arbeit wurden linguistische Eigenschaften deutscher und russischer wissenshaltiger Kontexte untersucht, Methoden der automatischen Extraktion wissenshaltiger Kontexte aus Textkorpora entwickelt und evaluiert und Perspektiven der praktischen Anwendung entsprechender Extraktionsergebnisse für die übersetzungsbezogene Terminologiearbeit diskutiert. Einleitend waren dabei die folgenden Fragen aufgeworfen worden:

- Wie wird fachspezifisches Wissen in Texten dargestellt, d. h. welche linguistischen Merkmale weisen wissenshaltige Kontexte auf?
- Welche Arten von fachspezifischem Wissen sind relevant für die übersetzungsorientierte Terminologiearbeit, d. h. welche Arten fachspezifischen Wissens sind in Terminologiedatenbanken darzustellen, um den Nutzern solcher Ressourcen einen effizienten und strukturierten, übersetzungsorientierten Wissenserwerb zu ermöglichen?
- Wie erkennen Übersetzer fachspezifisches Wissen in Texten, d. h. gibt es überhaupt einen Konsens in Bezug auf die Relevanz fachspezifischer Aussagen in Texten für die übersetzungsorientierte Terminologiearbeit?
- Wie kann relevantes semantisches Wissen in natürlichsprachigen Texten mit Hilfe von Computerprogrammen erkannt und automatisch aus ihnen extrahiert werden?
- Inwieweit sind automatisch aus Korpora extrahierte wissenshaltige Kontexte für die Terminologie überhaupt nutzbar?

Des Weiteren waren die folgenden Anforderungen an die durch vorliegende Arbeit zu erbringenden Erkenntnisse formuliert worden:

- Entwicklung, Implementierung und Evaluierung von Methoden für die Extraktion wissenshaltiger Kontexte für bislang nicht oder nur selten untersuchte, sich von den überwiegend bearbeiteten Sprachen Englisch und Französisch typologisch unterscheidende Sprachen. Konkret soll die Extraktion wissenshaltiger Kontexte für die Sprachen Deutsch und Russisch untersucht werden. Dies eröffnet nicht nur eine kontrastive Perspektive, sondern ermöglicht auch die Beurteilung existierender Extraktionsmethoden in Bezug auf ihre Leistungsfähigkeit unabhängig von einer bestimmten Sprache und leitet zu möglicherweise weiterführenden, multilingualen Arbeiten über.
- Die zu entwickelnden Methoden sollen für die Anwendung in unterschiedlichen Fachgebieten geeignet sein, d. h. es werden generische Methoden entwickelt. Für ihre Evaluierung werden daher Korpora verwendet, die unterschiedliche Fachgebiete abdecken. Der Nachteil einer solchen generischen Methode besteht freilich darin, dass sie den Besonderheiten stark spezialisierter Fachgebiete (z. B. Recht oder Medizin) unter Umständen nicht gerecht wird.
- Eine detaillierte Annotationsstudie und eine terminologieorientierte qualitative Bewertung von Extraktionsergebnissen sollen Aufschluss darüber geben, ob die Nutzer terminologischer Ressourcen – insbesondere Übersetzer – überhaupt einen Konsens in Bezug auf die Nützlichkeit bestimmter Aussagen entwickeln, und in welchem Maße automatisch extrahierte wissenshaltige Kontexte die terminologische Begriffsmodellierung unterstützen können.

Zur Untersuchung dieser breit angelegten Fragestellungen war zunächst in **Kapitel 2** der Begriff des wissenshaltigen Kontexts unter Bezugnahme auf die relevante linguistische Fachliteratur, aber auch auf stärker spezialisierte Arbeiten aus der Terminologiewissenschaft, der Philologie und der Computerlinguistik von anderen Kontexttypen sowie Elementen terminologischer Begriffsbeschreibungen abgegrenzt worden. Die in Kapitel 2 entwickelte eigene Definition des KRC-Begriffs führte einige zusätzliche Validitätskriterien für die Beschreibung empirisch auftretender wissenshaltiger Kontexte ein und bildete damit eine wichtige Grundlage der schließlich in Kapitel 4 beschriebenen Annotationsarbeiten, wobei von formalistischen oder an linguistischen Oberflächenkriterien ausgerichteten KRC-Klassifikationen zugunsten einer am Gehalt der tatsächlich durch einen KRC-Kandidaten vermittelten Information ausgerichteten Klassifikation Abstand genommen wurde. Mit welchen semantischen Informationstypen das so definierte Skelett des Begriffs auszufüllen sei, wurde schließlich in Form einer Typologie relevanter semantischer Relationen definiert.

Diese Darstellung antwortet in Teilen auf die oben wiederholte Frage Nr. 2 nach den Arten des durch terminologische Ressourcen zu modellierenden fachspezifischen bzw. semantischen Wissens und Anforderungen an die Darstellung desselben, wobei anzumerken ist, dass – auf diesen überaus wichtigen Punkt war bereits mehrfach hingewiesen worden – es sich hierbei um eine rein theoriebasierte Diskussion handelt, die empirischen, beispielsweise psycholinguistischen oder prozessorientierten Untersuchungen zu den Arten des für die Übertragung komplexer Fachtexte erforderlichen Wissens gegenüberzustellen und ggf. durch diese zu ergänzen oder gar zu revidieren ist. Aufgrund der Vielfalt und Unvorhersehbarkeit der an Übersetzer – auf eine Konzentration der Diskussion auf einen eng umgrenzten Themenbereich war einleitend ja gerade verzichtet worden (Anforderung Nr. 2), um übersetzerischen Arbeitsbedingungen Rechnung zu tragen – gestellten inhaltlich-fachlichen Anforderungen lässt sich das übersetzungsrelevante begriffliche Wissen nur schwer systematisieren, was auch die in Kapitel 4 beschriebene Annotationsstudie deutlich macht. Allerdings scheint zumindest die in Kapitel 2 geleistete Diskussion unterschiedlicher Relationstypologien, die immerhin gewisse Ähnlichkeiten zwischen den unterschiedlichen Typologien zutage förderte, darauf hinzudeuten, dass relevante Informationstypen als je nach Fachgebiet unterschiedlich gewichtete Konfigurationen von Qualia-Strukturen und umweltbezogenen (z. B. temporale und lokale Einordnung) Kategorien charakterisiert werden können. Im Hinblick auf die Anforderungen an die Darstellung komplexer Begriffsbeschreibungen in terminologischen Informationssystemen waren schließlich in Abschnitt 6.4 Desiderata für zukünftige Forschungsarbeiten formuliert worden, von denen das der Entwicklung individualisierbarer und dynamischer Darstellungsformen für den hier diskutierten Sachverhalt wohl als das wichtigste angesehen werden kann.

In **Kapitel 3** schließlich waren die wissenschaftlichen Grundlagen für die in dieser Arbeit zu leistende Untersuchung referiert worden. Konkret wurden sowohl wissenschaftliche Befunde zu linguistischen Eigenschaften definitorischer Aussagen als auch Annotations- und Extraktionsmethodologien diskutiert. Hierbei zeigte sich, dass nur wenige systematische Untersuchungen zu linguistischen Eigenschaften wissenshaltiger Kontexte oder Definitionen (insbesondere solche, die über die rein lexikalische Ebene hinausgehen) und kaum sprachübergreifende Studien existieren, was für die Relevanz vorliegender Arbeit spricht. Als am besten untersuchter Aspekt wurden in diesem Zusammenhang *knowledge patterns* identifiziert, für die in Bezug auf die Implementierung und die Komplexität der linguistischen Beschreibung jedoch eine erhebliche Variationsbreite beobachtet wurde. Im Hinblick auf die Annotation wissenshaltiger Kontexte in Textkorpora wurde zudem festgestellt, dass eine formalisierte Annotation nur wenigen der referierten Studien zugrunde liegt,

was insofern als problematisch angesehen werden kann, als die nachfolgend referierten Ansätze zur korpusbasierten Extraktion wissenshaltiger Kontexte und definitorischer Aussagen in Bezug auf ihre Performanz nicht miteinander verglichen werden können. Bisherige Erfahrungen mit der Annotation definitorischer Aussagen haben aber Hinweise auf erhebliche Schwierigkeiten erbracht, die sich auf die bereits genannten Faktoren – die schwierige konzeptuelle Fassbarkeit der Kategorie des wissenshaltigen Kontexts einerseits sowie die Unvorhersehbarkeit der inhaltlichen Anforderungen an die Nutzer terminologischer Ressourcen bei gleichzeitigem Fehlen belastbarer empirischer Studien andererseits – zurückführen lassen. Trotz der beeindruckenden Bandbreite bisher entwickelter Extraktionsmethodologien konnte darüber hinaus festgehalten werden, dass viele Extraktionsansätze nach wie vor an die Beschreibung – unterschiedlich formalisierter – sprachlicher Muster geknüpft sind, wobei allerdings häufig die – durch die im Verlauf dieser Arbeit angeführten Beispiele belegte – linguistische Vielgestaltigkeit empirisch beobachtbarer KRCs nicht berücksichtigt wird. Die Frage danach, ob die systematische Beschreibung linguistischer KRC-Eigenschaften Anstöße für die Verbesserung von Extraktionssystemen liefern kann, konnte mithin nicht nur aus praktischer Extraktionssicht, sondern auch im Hinblick auf die Fragestellung als relevant identifiziert werden, ob eine solche Beschreibung tatsächlich Anhaltspunkte für das in der referierten Literatur formulierte Postulat eines eigenen Aussagetypus „KRC“ ergibt, da Steigerungen der Extraktionsperformanz durch linguistisch informierte Ansätze nur dann zu erwarten sind, wenn diese Ansätze tatsächlich solche Merkmale nutzen, die wissenshaltige Kontexte von generischen Aussagen systematisch unterscheiden. Im Vergleich zu den in der Literatur häufig anzutreffenden vereinfachenden oder impressionistischen bzw. auf die reine Extraktion definitorischer Aussagen reduzierten Ansätzen wurde mithin eine wesentlich breiter angelegte Forschungsfrage formuliert.

In **Kapitel 4** wurde schließlich versucht, sowohl die Datengrundlage für die durchzuführende empirische Untersuchung herzustellen – und zwar in einer möglichst intersubjektiv verifizierbaren Art und Weise, um dem Problem der Nicht-Vergleichbarkeit von Studienergebnissen vorzubeugen und Schlussfolgerungen für zukünftige Annotationsexperimente formulieren zu können – als auch Hinweise darauf zu gewinnen, in welchem Maße Übersetzer überhaupt einen Konsens in Bezug auf den definitorischen Status von KRC-Kandidaten abzubilden vermögen (Frage und Anforderung Nr. 3). Die erstellten Goldstandard-Korpora sollten zudem unterschiedliche Fachgebiete abdecken (Anforderung Nr. 2). Im Ergebnis des Annotationsprozesses wurden Korpora eines mittleren Fachlichkeitsgrades erstellt, die unterschiedlichen, aber doch im weitesten Sinne „technischen“ Fachgebieten zugeordnet werden können. Schwierigkeiten bei der Annotation wissenshaltiger Kontexte ergaben sich dabei nicht nur aufgrund der oben genannten problematischen Aspekte, sondern auch aus dem Fehlen einer etablierten Annotationsmethodologie sowie aus der Spezifik der verwendeten Internet-Daten. Eine Besonderheit der hier durchgeführten Annotationsstudie bestand zudem darin, dass den Annotatoren keine differenzierten linguistischen Annotationsrichtlinien (wie etwa bei Walter (2010) oder Cramer (2011)) an die Hand gegeben worden waren, da ja gerade der zu erstellende Goldstandard als Grundlage für die die linguistische Beschreibung von KRCs dienen sollte, sondern vielmehr neben den eigens erarbeiteten, theoriebasierten Validitätskriterien die spezifische übersetzerische Kompetenz der Annotatoren den Ausschlag für Annotationsentscheidungen geben sollte.

Ein detailliertes und methodologisch stringentes Annotationsexperiment, aufgrund dessen Frage Nr. 3 vollumfänglich beantwortet werden könnte, wurde zwar nicht durchgeführt. Allerdings deuten die im Verlauf der Annotation gesammelten Daten darauf hin, dass die einzelnen Annotatoren unterschiedliche Annotationsstrategien ausgebildet haben, wodurch die Ausführungen Cramers (2011) zu Einflussfaktoren bei der Bewertung definitorischer Aussagen tendenziell bestätigt werden

können, da die beobachteten Unterschiede in der Annotationskonfidenz und der Akzeptanzquote der einzelnen Annotatoren nicht auf das bei den miteinander verglichenen Annotatorenpaaren identische Stimulusmaterial zurückgehen können. Aus dieser Beobachtung wurde der Schluss gezogen, dass im Zusammenhang mit zukünftigen Annotationsexperimenten Richtlinien nicht etwa in Bezug auf spezifische linguistische Strukturen, sondern im Hinblick beispielsweise auf den Umgang mit Zweifelsfällen formuliert werden sollten. Die Annotationsergebnisse wurden zudem – vorsichtig – dahingehend interpretiert, dass alle Annotatoren tatsächlich einen Grundstock an „Textsortenwissen“ bzw. eine prototypische Vorstellung in Bezug auf die sprachliche Form definitorischer Aussagen zu teilen scheinen. Aufbauend auf diesem Befund kann folglich davon ausgegangen werden, dass die in Kapitel 5 durchgeführte linguistische Analyse in der Tat linguistische Korrelate dieser Vorstellung offenlegt. Allerdings wurde auch argumentiert, dass das ebenfalls beobachtete erhebliche Maß an divergierenden Annotationsentscheidungen die wesentliche Rolle der von Cramer (2011) genannten Einflussfaktoren bestätigt, woraus sich die Notwendigkeit weiterführender und umfassenderer Studien ergibt.

Kapitel 5 war der Untersuchung linguistischer Merkmale wissenshaltiger Kontexte (Frage Nr. 1) mit Hilfe korpuslinguistischer Methoden vorbehalten. Ein Katalog interessierender sprachlicher Merkmale war dabei theoriebasiert durch Kombination mehrerer interdisziplinärer Beschreibungsansätze erarbeitet worden. Im Ergebnis dieser Untersuchungen konnte gezeigt werden, dass deutsche und russische KRCs – zumindest in den untersuchten Daten – durchaus durch spezifische Konfigurationen linguistischer Merkmale insbesondere verallgemeinernder und depersonalisierender Natur beschrieben werden können. Allerdings wurde auch festgehalten, dass statt von einem eigenen Funktionalstil des Erklärens und Definierens gerade aufgrund der Generizität bzw. linguistischen Unauffälligkeit von vielen der beobachteten „KRC-Merkmale“, aufgrund derer eine vollständige Abgrenzung wissenshaltiger Kontexte von generischen Aussagen kaum möglich scheint, eher von Kombinationen bestimmter Merkmalsgruppen gesprochen werden sollte, die auf das Vorliegen eines wissenshaltigen Kontexts hindeuten. Andererseits scheinen die Untersuchungsergebnisse zumindest gewisse Gemeinsamkeiten zwischen russischen und deutschen KRCs zu belegen, was die Hypothese einer wenigstens teilweisen funktionalen Determination des Aussagetypus‘ „KRC“ über Sprachengrenzen hinweg stützt.

In **Kapitel 6** wurden schließlich praktische Extraktionsmethodologien erprobt und damit Hinweise für die Beantwortung der Frage Nr. 4 im Sinne von Anforderung Nr. 1 erarbeitet. Konkret wurden zunächst für das Russische und das Deutsche musterbasierte Extraktionssysteme entwickelt, wobei als *knowledge patterns* Verben in einem syntaxbasierten *setting* herangezogen wurden, die bereits im Ergebnis der in Kapitel 5 durchgeführten linguistischen Analysen identifiziert worden waren. Die Extraktionsergebnisse bestätigten insofern die bereits dort formulierten Zweifel in Bezug auf die diskriminatorische Kraft solcher Muster, als die Übertragung von Extraktionsmustern auf ungesehene Daten nur in begrenzten Maße gelang. Dieses Ergebnis scheint zudem dadurch gestützt zu werden, dass auch die theoriebasierte Erweiterung der Mustermenge für das Deutsche nur eine geringfügige Performanzverbesserung erbachte, die in der Literatur beschriebenen Muster sich mithin ebenfalls als weitgehend korpusspezifisch erwiesen haben. Allerdings konnte für das Russische – bei einer zugegebenermaßen weitaus niedrigeren Ausgangsperformanz – durch die theoriebasierte Spezifikation zusätzlicher Muster eine deutliche Leistungssteigerung erreicht werden. Geschlussfolgert worden war aus diesen Ergebnissen, dass linguistische Extraktionsmuster möglichst in einer mathematisch formalisierten Art und Weise auf größeren als den hier zugrunde gelegten Datenmengen erlernt werden sollten, und zwar nicht nur wegen des mit der manuellen Musterspezifikation verbundenen Arbeitsaufwandes, sondern auch aufgrund der kaum

kalkulierbaren Leistungsfähigkeit von Musterkatalogen, die auf kleinen Datensätzen beruhen. Ob freilich automatisch erlernte Musterkataloge die Anforderung der allgemeinen Verwendbarkeit (Anforderung Nr. 2) bzw. die Hearstschen Qualitätskriterien besser erfüllen als die in vorliegender Arbeit genutzten Extraktionsmuster ist dabei eine noch zu beantwortende Frage. Der syntaxbasierte Ansatz der KRC-Extraktion scheint sich dagegen – wenn einmal von den nun hinlänglich diskutierten Limitationen musterbasierter Ansätze generell sowie dem Problem der Bindung an entsprechende Annotationswerkzeuge ausreichender Güte einmal abgesehen wird – sowohl für das Deutsche als auch für das Russische zu eignen.

Im Anschluss an die symbolisch ausgerichtete Untersuchung waren auch hybride Extraktionsmethoden angewandt worden, wobei die bereits in Kapitel 5 erprobte Methode der logistischen Regression Anwendung fand. Als vielversprechendster Ansatz wurde dabei der Einbezug musterbezogener Informationen in ein ansonsten auf einfachen quantitativen Merkmalen (den KRC-Merkmalen aus Kapitel 5) basierendes System identifiziert. Für die KRC-Identifikation in beiden Sprachen erwiesen sich neben musterbezogenen Informationen insbesondere lexikalische, terminologische und grammatische Informationstypen bei interlingual leicht unterschiedlicher Gewichtung als wesentlich, wobei auch hier wieder die bereits in Kapitel 5 herausgearbeitete Tendenz zur Verwendung unmarkierter linguistischer Strukturen bestätigt werden konnte. Nicht zuletzt konnte im Ergebnis der durchgeführten Experimente festgehalten werden, das die durch die Kombination musterbezogener Informationen mit den in Kapitel 5 erhobenen KRC-Merkmalen beobachtete Leistungssteigerung für die Relevanz und Validität dieser Merkmale spricht, wobei auch – obgleich die Mehrheit der zumindest hier analysierten Daten anhand eher simpler linguistischer Strukturen beschrieben werden kann – linguistische Strukturen in den Blick kamen, die in der relevanten Literatur nur wenig Beachtung finden (z. B. die Problematik des russischen Verbalaspekts oder subjektloser Satzstrukturen). Insgesamt scheint ein hybrider Ansatz der KRC-Extraktion wie der in vorliegender Arbeit entwickelte dazu beizutragen, die inhärente Leistungsgrenze musterbasierter Methoden ein wenig in Richtung einer besseren Gesamtperformanz zu verschieben. Für weitergehende kontrastive Studien zu wissenshaltigen Kontexten hat vorliegende Arbeit damit einige interessante Anstöße gegeben.

In Bezug auf die Nutzbarkeit von automatisch aus Textkorpora extrahierten Daten für die praktische Terminologiearbeit war allerdings festgestellt worden, dass einer Nutzung in der terminologischen Praxis momentan nicht nur technische, sondern auch konzeptuelle Hürden entgegenstehen. Neben eher technisch-computerlinguistischen Problemen, die vor einer direkten Integration von Extraktionsergebnissen in terminologische Ressourcen noch gelöst werden müssen, waren hier auch Fragestellungen in Bezug auf die diskursive Einbindung von Begriffsbeschreibungen in größere Textzusammenhänge und die Repräsentation umfangreicher Begriffsbeschreibungen in terminologischen Datensammlungen genannt worden. Nicht zuletzt scheint das Problem der allgemeinen Qualitätssicherung nach wie vor umfangreiche manuelle Interventionen vorauszusetzen.

7.2 Quo vadis, korpusbasierte Terminologie?

Zum Abschluss der in dieser Arbeit durchgeführten Untersuchung ist noch zu diskutieren, inwieweit die einleitend aufgezeigten Perspektiven der Beschäftigung mit aus Textkorpora automatisch extrahierten KRCs bzw. generell mit korpusbasierten Ansätzen für die wissenschaftliche und praktische Beschäftigung mit terminologischen Fragestellungen angesichts der erarbeiteten Ergebnisse weiter spezifiziert werden können oder müssen. Hervorzuheben sind in diesem Zusammenhang aus einer terminologiepraktischen Sicht insbesondere die Aspekte Qualität und

Einbindung in terminologische *workflows*, wohingegen aus einer wissenschaftlichen Perspektive in erster Linie die Notwendigkeit einer stärker empirischen Orientierung in den Blick kommt.

7.2.1 Terminologiepraktische Aspekte

Aus einem praxisorientierten Blickwinkel war in der Einleitung zu dieser Arbeit argumentiert worden, dass die Beschäftigung mit korpusbasierten Methoden der Informationsextraktion wie beispielsweise der Gewinnung wissenshaltiger Kontexte nicht nur Wege hin zu einer effizienten und zeitsparenden Erstellung umfangreicherer, informativerer und damit für die Nutzer sinnvoller verwendbarer terminologischer Ressourcen eröffnet, sondern auch Ansatzpunkte für die Nutzung semantischer Technologien wie etwa einer relationenbasierten Suche bietet. Die in den vorstehenden Kapiteln dargestellte Untersuchung hat nun ergeben, dass diese Perspektive nach wie vor besteht, ihre Umsetzung jedoch an einige Voraussetzungen geknüpft ist.
Als wesentliches Problem erweist sich dabei zunächst das der Qualität terminologischer Datensammlungen: Während für einige computerlinguistische Teildisziplinen gegenwärtig das *big-data*-Paradigma, demzufolge ein Mehr an Daten quasi automatisch auch ein Mehr an Qualität bedeutet, als dominant angesehen werden kann, hat sich im Zuge dieser Arbeit gezeigt, dass aus terminologischer Sicht bereits bei der Analyse kleiner Datenmengen verstärkt auf dateninterne Qualitätsaspekte zu achten ist. Gemeint ist damit nicht nur, dass Korpora, auch wenn sie *ad hoc* unter Nutzung des Internets erstellt werden, gewissen Qualitätsanforderungen genügen (und beispielsweise überwiegend Fachtexte im engeren Sinne enthalten) oder dass leistungsfähigere Extraktionsmethoden entwickelt werden sollten, sondern auch, dass vollumfängliche Begriffsbeschreibungen im hier verstandenen Sinne die Anwendung elaborierter Filtertechniken sowohl während der Extraktion relevanter Informationen als auch im Hinblick auf die Präsentation der erarbeiteten Daten voraussetzen. Das in der Einleitung zu dieser Arbeit angesprochene korpuslinguistische Relevanzprinzip, demzufolge relevante und korrekte Daten bereits aufgrund ihrer schieren Auftretenshäufigkeit von irrelevanten oder fehlerhaften Daten unterschieden werden können, wird dadurch nicht völlig ausgehebelt. Es zeigt sich aber, dass in der Terminologie auch weniger häufig auftretende Benennungen oder Beschreibungen von Interesse sein können, nämlich dann, wenn es sich tatsächlich um hochspezialisierte Wissensbausteine handelt. Vor diesem Hintergrund haben korpusorientierte terminologische Forschungen bislang kaum zur Etablierung eines an quantitativen oder eben korpusbasierten Kriterien orientierten Qualitätsbegriffs beigetragen, während in der klassischen Terminologielehre unter Qualität im Wesentlichen Standardkonformität bzw. das Ergebnis des bewussten Einsatzes der fachlichen Fähigkeiten von Einzelpersonen verstanden zu werden scheint, folglich ein nur bedingt operationalisier- und messbarer Qualitätsbegriff vorliegt. Nicht zuletzt aus diesen Gründen werden manuelle Qualitätssicherungsmaßnahmen auch in solchen *settings*, in denen korpusbasierte Methoden zum Einsatz kommen sollen, vorerst unverzichtbar bleiben. Angesichts des Potentials korpusbasierter Methoden sollten allerdings Qualitätskriterien, Standards oder zumindest konsolidierte Prozeduren und Prozesse im Sinne von *best practices* entwickelt werden, welche die Messung der Qualität beispielsweise von Extraktionsergebnissen oder Datensammlungen ebenso ermöglichen wie stringente Entscheidungsfindungen. In dieser Hinsicht könnte die Terminologie sogar insofern eine Vorreiterrolle vor anderen texttechnologisch orientierten Disziplinen einnehmen, als sie das möglicherweise etwas naive *big-data*-Paradigma durch differenzierte Standpunkte und Methoden ergänzen kann. Kann dagegen das Fach keine entsprechenden Fortschritte machen, besteht Grund zu

der Befürchtung, dass korpusbasierte Verfahren in der praktischen Terminologiearbeit auch auf längere Sicht nur in rudimentärer Form zum Einsatz kommen werden.
Eng mit dieser Problematik verbunden ist auch die Frage nach sinnvollen Möglichkeiten der Einbindung korpusbasierter Verfahren in terminologische *workflows*. Während viele terminologische und generell übersetzerische Arbeitstechniken wie beispielsweise die Recherche in Paralleltexten als vom Ansatz her korpuslinguistisch charakterisiert werden können, ist für das Gros der gegenwärtig aktiven Übersetzer und Terminologen nach wie vor von einer unzureichenden Vertrautheit mit korpuslinguistischen Methoden und Werkzeugen auszugehen. Hieraus ergibt sich nicht nur eine didaktische Aufgabe im Hinblick auf die Ausbildung des professionellen Nachwuchses, sondern selbstverständlich auch die Notwendigkeit zu eruieren, auf welche Weise (beispielsweise der technische und zeitliche) Aufwand und Nutzen (ein Mehr an Informativität sowie erweiterte Möglichkeiten der Personalisierung und Abfrage, Zeiteinsparung bei Vorliegen hochperformanter Systeme) miteinander in Einklang zu bringen sind. Für zukünftige terminologische Forschungen ergibt sich hiermit ein überaus praxisrelevantes Arbeitsfeld, das über die von Roelcke (2013) kritisierte reine Beschreibung des lexikalischen Inventars einzelner Fachsprachen deutlich hinausgeht.

7.2.2 Terminologiewissenschaftliche Aspekte

Das Potential der Beschäftigung mit korpusbasierten Ansätzen wie etwa der KRC-Extraktion aus Textkorpora für die Terminologiewissenschaft wurde in der Einleitung zu dieser Arbeit daran festgemacht, dass entsprechende Ansätze es ermöglichen, begriffliche Systematisierungen an empirische Daten beispielsweise quantitativer Natur zu knüpfen. Begriffliche Festlegungen wären dann an quantifizierbare, mit Hilfe etablierter korpuslinguistischer Methoden erhobene Daten gebunden und damit zumindest falsifizierbar, wobei natürlich auch in dieser Hinsicht wieder auf das Fehlen stringent anwendbarer Entscheidungskriterien bzw. die Notwendigkeit der Erarbeitung entsprechender Kriterien hinzuweisen ist. Verwiesen wurde des Weiteren darauf, dass auf Grundlage von Untersuchungen, in deren Zuge beispielsweise die genannten Entscheidungskriterien erarbeitet werden könnten, auch alltägliche Arbeitspraktiken von Terminologen und Übersetzern, nämlich im Wesentlichen Techniken der textbasierten Recherche terminologischer Information, einer systematischen wissenschaftlichen Beschreibung zugänglich gemacht werden könnten und damit didaktischen sowie weiterführenden Zwecken zur Verfügung stünden.
Tatsächlich wurde in dieser Arbeit immer wieder auf die Notwendigkeit nicht nur umfangreicherer Untersuchungen zum Zwecke der Replikation bzw. Ergänzung und Differenzierung der hier dargestellten Ergebnisse, sondern auch nutzerorientierter Studien generell hingewiesen. Gemeint ist damit im Wesentlichen, dass sich die Terminologiewissenschaft von einer auf die (je nach Szenario mehr oder weniger normative) Beschreibung fachsprachlicher Lexika orientierten Disziplin in Richtung eines Fachs entwickeln sollte, dass sich mit der Kommunikation bzw. dem Transfer spezialisierten Wissens mit den Mitteln der Sprache generell beschäftigt. Fachspezifische Termini sind in zahlreiche kognitive, kommunikative und diskursive Zusammenhänge eingebunden, allein terminologische Forschungen haben diesem Umstand ungeachtet der mitunter recht polemisch geführten fachinternen Diskussion bislang nur in begrenztem Maße Rechnung getragen. Die kognitive Dimension der Terminologie, auch hierauf wurde mehrfach hingewiesen, wurde bislang kaum berücksichtigt, so dass sich dieses Versäumnis – dies zeigen auch die in dieser Arbeit diskutierten Schwierigkeiten der theoretisch-methodologischen Konzeptualisierung wissenshaltiger Kontexte – regelmäßig in Modellierungsproblemen niederschlägt. Die in der Einleitung zitierte pointierte Kritik Gruntar Jermols (2010) kommt mithin nicht von ungefähr. Im Hinblick auf den

Beitrag, den die Beschäftigung mit wissenshaltigen Kontexten bezüglich dieser Aspekte leisten kann, ist festzuhalten, dass KRCs überhaupt erst einmal die Erstellung einer empirischen Datenbasis zum Gebrauch und zur Definition fachspezifischer Termini durch die Nutzer der Terminologie selbst ermöglichen und damit Hinweise auf die relevanten kognitiven Strukturen der Verfasser von Fachtexten geben.

Bibliographie

Agichtein, Eugene / Gravano, Luis (2000): „*Snowball*: Extracting Relations from Large Plain-Text Collections“. Fifth ACM International Conference on Digital Libraries, 2.-7. Juni 2000, San Antonio, USA, 85-94, http://www.mathcs.emory.edu/~eugene/papers/dl00.pdf, abgerufen am 25.08.2014.

Ahmad, Khurshid (2001): „The Role of Specialist Terminology in Artificial Intelligence and Knowledge Acquisition”. Wright, Sue Ellen / Budin, Gerhard (Hg.): *Handbook of Terminology Management. Bd. 2: Application-Oriented Terminology Management*. Amsterdam/Philadelphia: John Benjamins, 809-844.

Ahmad, Khurshid / Rogers, Margaret (2001): „Corpus Linguistics and Terminology Extraction”. Wright / Budin (Hg.): *Handbook of Terminology Management 2*, 725-760.

Alshawi, Hiyan (1989): „Analysing the dictionary definitions“. Boguraev, Bran / Briscoe, Ted (Hg.): *Computational Lexicography for Natural Language Processing.* London/New York: Longman, 153-169.

Altmann, Gabriel (1988): *Wiederholungen in Texten.* (Quantitative Linguistics 36). Bochum: Brockmeyer.

Anthony, Laurence (2005): „AntConc. A Learner and Classrom Friendly, Multi-Platform Corpus Analysis Toolkit”. IWLeL 2004: An Interactive Workshop on Language e-Learning, 10. Dezember 2004, Tokio, Japan, 7-13, http://www.antlab.sci.waseda.ac.jp/research/iwlel_2004_anthony_antconc.pdf abgerufen am 25.08.2014.

Arnauld A. / Nicole P. (1996): *Logic or the Art of Thinking*. Cambridge: Cambridge University Press.

Artstein, Ron / Poesio, Massimo (2008): „Inter-Coder Agreement for Computational Linguistics”. Computational Linguistics 34 (4), 555-596.

Atkins, Sue / Clear, Jeremy / Ostler, Nicholas (1992): „Corpus Design Criteria”. Literary and Linguistic Computing 7 (1), 1-16.

Auger, Alain / Barrière, Caroline (2008): „Pattern-based approaches to semantic relation extraction”. Terminology 14 (1), 1-19.

Aussenac-Gilles, Nathalie / Biébow, Brigitte / Szulman, Sylvie (2000): „Revisiting Ontology Design: A Method Based on Corpus Analysis”. Dieng, Rose / Corby, Olivier (Hg.): *Knowledge Engineering and Knowledge Management. Methods, Models, and Tools.* (Lecture Notes in Computer Science 1937). Berlin/Heidelberg: Springer, 172-188.

Aussenac-Gilles, Nathalie / Jacques, Marie-Paule (2008): „Designing and evaluating patterns for relation acquisition from text with CAMÉLÉON”. Terminology 14 (1), 45-73.

Baker, Paul / Hardie, Andrew / McEnery, Tony (2006): *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baranov, Oleg S. (1995): *Ideografičeskij slovar’ russkogo jazyka* (= Ideographisches Wörterbuch der russischen Sprache). Moskva: ĖTS.

Baroni, Marco / Bernardini, Silvia (2004): „Bootcat: Bootstrapping Corpora and Terms from the Web”. LREC 2004, 24.-30. Mai 2004, Lissabon, Portugal, 1313-1316, http://www.lrec-conf.org/proceedings/lrec2004/, abgerufen am 25.08.2014.

Baroni, Marco / Evert, Stefan (2009): „Statistical methods for corpus exploitation”. Lüdeling, Anke / Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. *Bd. 2*. (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2). Berlin/New York: de Gruyter, 777-803.

Barnbrook, Geoff (2002): *Defining Language. A local grammar of definition sentences.* (Studies in Corpus Linguistics 11). Amsterdam/Philadelphia: John Benjamins.

Barrière, Caroline (2004): „Knowledge-rich Contexts Discovery”. Tawfik, Ahmed Y. / Goodwin, Scott D. (Hg.): *Advances in Artificial Intelligence.* (Lecture Notes in Computer Science 3060). Berlin/Heidelberg: Springer, 187-201.

Barrière, Caroline (2010): „Expanding terminological knowledge using web corpora". Präsentation auf der Konferenz „Approaches to the Lexicon", 8.-9. Dezember 2010, Kopenhagen, Dänemark, https://conference.cbs.dk/index.php/lexicon/lexicon/paper/viewFile/837/494, abgerufen am 15.05.2014.

Baumann, Klaus-Dieter (1998): „Fachsprachliche Phänomene in den verschiedenen Sorten von populärwissenschaftlichen Vermittlungstexten". Hoffmann, Lothar / Kalverkämper, Hartwig / Wiegand, Herbert Ernst / Galinski, Christian / Hüllen, Werner (Hg.): *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. Bd. 1.* (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1). Berlin/New York: de Gruyter, 728-735.

Bean, Carol A. / Green, Rebecca (2001) (Hg.): *Relationships in the Organization of Knowledge.* (Information Science and Knowledge Management 2). Dordrecht: Kluwer.

Bendixen, Bernd / Voigt, Wolfgang (1999): *Russische Grammatik. Ein Leitfaden.* Wiesbaden: Harrassowitz.

Ben Abacha, Asma / Zweigenbaum, Pierre (2011): „Automatic Extraction of Semantic Relations between Medical Entities: a Rule Based Approach". Journal of Biomedical Semantics 2 (5), S4.

Beghtol, Clare (2001): „Relationships in Classificatory Structure and Meaning". Bean / Green (Hg.): *Relationships in the Organization of Knowledge*, 99-113.

Bierwisch, Manfred / Kiefer, Ferenc (1969): „Remarks on Definitions in Natural Language". Kiefer, Ferenc (Hg.): Studies in Syntax and Semantics. (Foundations of Language 10). Dordrecht: Reidel, 55-79.

Biébow, Brigitte / Szulman, Sylvie (1999): „TERMINAE: A Linguistics-Based Tool for the Building of a Domain Ontology". Studer, Rudi (Hg.): *Knowledge Acquisition, Modeling and Management.* (Lecture Notes in Computer Science 1621). Berlin/Heidelberg: Springer, 49-66.

Björkelund, Anders / Bohnet, Bernd / Hafdell, Love / Nugues, Pierre (2010): „A high-performance syntactic and semantic dependency parser". COLING 2010, 23.-27. August, Peking, China, Demonstration Volume, 33-36, http://www.aclweb.org/anthology/C10-3009, angerufen am 25.08.2014.

Blohm, Sebastian (2010): *Large-Scale Pattern-Based Relation Extraction from the World Wide Web.* Dissertation. Karlsruhe: Karlsruher Institut für Technologie.

Boguraev, Branimir / Pustejovsky, James (1996) (Hg.): *Corpus Processing for Lexical Acquisition.* (Language, Speech and Communication). Cambridge/London: MIT Press.

Bohnet, Bernd (2010): „Very High Accuracy and Fast Dependency Parsing is not a Contradiction". COLING 2010, 23.-27. August 2010, Peking, China, 89-97, http://www.aclweb.org/anthology/C10-1011, abgerufen am 25.08.2014.

Borg, Claudia (2009): *Automatic Definition Extraction Using Evolutionary Algorithms.* Master-Arbeit. Valletta: Universität Malta.

Bourigault, Didier / Slodzian, Monique (1999): „Pour une terminologie textuelle". Terminologies nouvelles 19, 20-32.

Bowker, Lynne / Pearson, Jennifer (2002): *Working with Specialized Language. A Practical Guide to Using Corpora.* London: Routledge.

Brants, Sabine / Dipper, Stefanie / Hansen, Silvia / Lezius, Wolfgang / Smith, George (2002): „The TIGER Treebank". http://www.coli.uni-saarland.de/publikationen/softcopies/Brants:2002:TT.pdf, abgerufen am 15.05.2014.

Buchholz, Sabine / Marsi, Erwin (2006): „CoNLL-X shared task on Multilingual Dependency Parsing." CoNLL-X 2006, 8.-9. Juni 2006, New York, USA, 149-164, http://ilk.uvt.nl/~emarsi/download/pubs/14964.pdf, abgerufen am 25.08.2014.

Budin, Gerhard / Bühler, Hildegund (1999): „Grundsätze und Methoden der neueren Terminographie". Hoffmann, Lothar / Kalverkämper, Hartwig / Wiegand, Herbert Ernst / Galinski, Christian / Hüllen, Werner (Hg.): *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft*. Bd. 2. (Handbücher zur Sprach- und Kommunikationswissenschaft 14.2). Berlin/New York: de Gruyter, 2096-2108.

Bußmann, Hadumod (2002[3]) (Hg.): *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner.

Cimiano, Philipp (2006): *Ontology learning and population from text: algorithms, evaluation and application*. New York: Springer.

Condamines, Anne (2002): „Corpus analysis and conceptual relation patterns". Terminology 8 (1), 141-162.

Condamines, Anne / Rebeyrolle, Josette (1998): „CTKB: A corpus-based approach to a Terminological Knowledge Base". Computerm 1998, COLING-ACL 1998, August 1998, Montréal, Kanada, 29-35, http://w3.erss.univ-tlse2.fr/membres/rebeyrol/Articles/condamines_rebeyrolle_1998.pdf, abgerufen am 22.03.2011.

Condamines, Anne / Rebeyrolle, Josette (2001): „Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB)". Bourigault, Didier / Jacquemin, Christian, L'Homme, Marie-Claude (Hg.): *Recent Advances in Computational Terminology*. (Natural Language Processing 2). Amsterdam/Philadelphia: John Benjamins, 127-148.

Cramer, Irene Magdalena (2011): *Definitionen in Wörterbuch und Text: Zur manuellen Annotation, korpusgestützten Analyse und automatischen Extraktion definitorischer Textsegmente im Kontext der computergestützten Lexikographie*. Dissertation. Dortmund: TU Dortmund.

Cruse, David Alan (2000): *Meaning in Language. An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

Cunningham, Hamish / Maynard, Diana / Bontcheva, Kalina / Tablan, Valentin / Aswani, Niraj / Roberts, Ian / Gorrell, Genevieve / Funk, Adam / Roberts, Angus / Damljanovic, Danica / Heitz, Thomas / Greenwood, Mark A. / Saggion, Horacio / Petrak, Johann / Li, Yaoyong / Peters, Wim (2011): *Text Processing with GATE. Version 6.* http://gate.ac.uk/sale/tao/split.html, abgerufen am 28.03.2014.

de Groc, Clément (2011): „Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction". IEEE/WIC/ACM: International Conference on Web Intelligence and Intelligent Agent Technology, 22.-27. August 2011, Lyon, Frankreich, 497-498, http://dbonline.igroupnet.com/ACM.Ft/2060000/2052397/4513a497.pdf, abgerufen am 25.08.2014.

de Saussure, Ferdinand (1966): *Course in General Linguistics*. New York: McGraw-Hill (französische Erstausgabe 1916).

Del Gaudio, Rosa / Branco, António (2007): „Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach". Neves, José / Santos, Manuel Filipe / Machado, José Manuel (Hg.): *Progress in Artificial Intelligence.* (Lecture Notes in Artificial Intelligence 4874). Berlin/Heidelberg: Springer, 659-670.

Dornseiff, Franz (2004[8]): *Der deutsche Wortschatz nach Sachgruppen.* Berlin/New York: de Gruyter.

Drescher, Martina (1992): *Verallgemeinerung als Verfahren der Textkonstitution. Untersuchungen zu französischen Texten aus mündlicher und schriftlicher Kommunikation.* Stuttgart: Steiner.

Dubuc, Robert / Lauriston, Andy (1997): „Terms and Contexts". Wright, Sue Ellen / Budin, Gerhard (Hg.): *Handbook of Terminology Management*. *Bd. 1: Basic Aspects of Terminology Management*. Amsterdam/Philadelphia: John Benjamins, 80-87.

Engelberg, Stefan / Lemnitzer, Lothar (2008[3]): *Lexikographie und Wörterbuchbenutzung.* (Stauffenburg Einführungen 14). Tübingen: Narr.

Engwall, Gunnel (1994): „Not Chance but Choice: Criteria in Corpus Creation". Atkins, Sue / Zampolli, Antonio (Hg.): *Computational Approaches to the Lexicon.* Oxford: Oxford University Press, 49-82.

Erjavec, Tomaž (2010): „MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora". LREC 2010, 17.-23. Mai 2010, Valletta, Malta, 2544-2547, http://www.lrec-conf.org/proceedings/lrec2010/index.html, abgerufen am 25.08.2014.

Evert, Stefan (2009): „Corpora and collocations". Lüdeling / Kytö (Hg.): *Corpus Linguistics 2*, 1212-1248.

Evert, Stefan / Baroni, Marco (2007): „zipfR: Word Frequency Distributions in R". ACL 2007, 23.-30. Juni 2007, Prag, Tschechische Republik, 29-32, http://www.stefan-evert.de/PUB/EvertBaroni2007.pdf, abgerufen am 25.08.2014.

Fahmi, Ismail / Bouma, Gosse (2006): „Learning to Identify Definitions using Syntactic Features". Workshop on Learning Structured Information in Natural Language Applications, EACL 2006, 3. April 2006, Trento, Italien, 64-71, http://ai-nlp.info.uniroma2.it/eacl2006-ws10/WS10-eacl2006-proceedings.pdf, abgerufen am 25.08.2014.

Feliu, Judit / Cabré, Maria Teresa (2002): „Conceptual relations in specialized texts: new typology and an extraction system proposal". TKE 2002, 28.-30. August 2002, Nancy, Frankreich, 45-49.

Fellbaum, Christiane (Hg.) (1998): *WordNet. An Electronic Lexical Database.* Cambridge: MIT Press.

Fišer, Darja (2009): *Izdelava slovenskega semantičnega leksikona z uporabo eno- in večjezičnih jezikovnih virov.* (= Erstellung eines slowenischen semantischen Lexikons mit Hilfe ein- und mehrsprachiger Ressourcen). Dissertation. Ljubljana: Universität Ljubljana.

Fišer, Darja / Pollak, Senja / Vintar, Špela (2010): „Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources". LREC 2010, 17.-23. Mai 2010, Valletta, Malta, 2932-2936, http://www.lrec-conf.org/proceedings/lrec2010/index.html, abgerufen am 25.08.2014.

Fletcher, William H. (2004): „Making the Web More Useful as a Source for Linguistic Corpora". Connor, Ulla / Upton, Thomas A. (Hg.): *Applied Corpus Linguistics. A Multidimensional Perspective.* Amsterdam/New York: Rodopi, 191-205.

Geeraerts, Dirk (2010): *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Gerzymisch-Arbogast, Heidrun (1996): *Termini im Kontext*. *Verfahren zur Erschließung und Übersetzung der textspezifischen Bedeutung von fachlichen Ausdrücken*. (Forum für Fachsprachenforschung 31). Tübingen: Narr.

Girju, Roxana / Nakov, Preslav / Nastase, Vivi / Szpakowicz, Stan / Turney, Peter D. / Yuret, Deniz (2007): „SemEval-2007 Task 04: Classification of Semantic Relations between Nominals". 4th International Workshop on Semantic Evaluations, ACL 2007, 23.-24. Juni 2007, Prag, Tschechische Republik, 13-18, http://people.ischool.berkeley.edu/~nakov//selected_papers_list/semeval2007_task4.pdf, abgerufen am 25.08.2014.

Göpferich, Susanne (1995): *Textsorten in Naturwissenschaften und Technik: pragmatische Typologie - Kontrastierung - Translation.* (Forum für Fachsprachen-Forschung 27). Tübingen: Narr.

Graf, Ulrich / Henning, Hans-Joachim / Stange, Kurt / Wilrich, Peter-Theodor (1998[3]): *Formeln und Tabellen der angewandten mathematischen Statistik.* Berlin: Springer.

Green, Rebecca / Bean, Carol A. / Myaeng, Sung H. (2002): *The Semantics of Relationships: An Interdisciplinary Perspective.* (Information Science and Knowledge Management 3). Dordrecht: Springer.

Gries, Stefan (2005): „Null-hypothesis significance testing of word-frequencies: a follow-up on Kilgarriff". Corpus Linguistics and Linguistic Theory 1(2), 277-294.

Gries, Stefan (2009): *Statistics for Linguistics with R. A Practical Introduction.* Berlin: de Gruyter.

Gruntar Jermol, Ada (2010): „Recht für Juristen, Recht für Laien? – Oder: (Wie) Lässt sich das Recht dem juristischen Laien näherbringen?". Lebende Sprachen 55 (1), 31-49.

Halliday, Michael A. K. (1994[2]): *An introduction to functional grammar.* London: Arnold.

Halskov, Jakob /Barrière, Caroline (2008): „Web-based extraction of semantic relation instances for terminology work". Terminology 14 (1), 20-44.

Hamp, Birgit / Feldweg, Helmut (1997): „GermaNet – a Lexical-Semantic Net for German". Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, ACL/EACL 1997, 12. Juli 1997, Madrid, Spanien, 9-15, http://www.aclweb.org/anthology/W/W97/W97-0802.pdf, abgerufen am 25.08.2014.

Hanks, Patrick (2006a): „Context, Communicative". Brown, Keith (Hg.): *Encyclopedia of Language and Linguistics*. Amsterdam: Elsevier, 115-128.

Hanks, Patrick (2006b): „Definition". Brown (Hg.): *Encyclopedia of Language and Linguistics*, 399-402.

Hearst, Marti A. (1992): „Automatic Acquisition of Hyponyms from Large Text Corpora". COLING 1992, 23.-28. August 1992, Nantes, Frankreich, 539-545, http://people.ischool.berkeley.edu/~hearst/papers/coling92.pdf, abgerufen am 25.08.2014.

Helbig, Hermann (2001): *Die semantische Natur natürlicher Sprache*. *Wissensrepräsentation mit MultiNet*. Berlin/Heidelberg: Springer.

Herbst, Thomas / Klotz, Michael (2003): *Lexikografie*. Paderborn: Schöningh.

Hoffmann, Ludger (2009): „Adverb". Hoffmann, Ludger (Hg.): *Handbuch der deutschen Wortarten.* Berlin/New York: de Gruyter, 223-264.

Hohnhold, Ingo (1999): „Übersetzungsorientierte Terminographie: Grundsätze und Methoden". Hoffmann / Kalverkämper / Wiegand with Galinski / Hüllen (Hg.): *Fachsprachen.* Bd. 2, 2155-2164.

Holtz, Monica / Teich, Elke (2009): „Design of the Darmstadt Scientific Text Corpus". http://www.linglit.tu-darmstadt.de/fileadmin/linglit/holtz/DaSciTex/dfg-corpus-design.pdf, abgerufen am 05.05.2014.

Hundt, Marianne / Sand, Andrea / Siemund, Rainer (1999): „Manual of Information to accompany The Freiburg – LOB Corpus of British English ('FLOB'). http://khnt.aksis.uib.no/icame/manuals/flob/INDEX.HTM, abgerufen am 09.01.2014.

International Organization for Standardization (2000): *International Standard ISO 1087-1: 2000 – Terminology Work – Vocabulary – Part 1: Theory and application*. Genf: ISO.

International Organization for Standardization (2009a): *International Standard ISO 12620: 2009 – Terminology and Other Language and Content Resources – Specification of Data Categories and Management of a Data Category Registry for Language Resources*. Genf: ISO.

International Organization for Standardization (2009b): *International Standard ISO 704: 2009 – Terminology Work – Principles and Methods.* Genf: ISO.

Iomdin, L. / Petrochenkov, V. / Sizov, V. / Tsinman, L. (2012): „ETAP Parser: State of the Art". Dialogue 2012, 30. Mai - 3. Juni 2012, Bekasovo, Russland, 119-131, http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Iomdin.pdf, abgerufen am 25.08.2014.

Jin, Yiping / Kan, Min-Yen / Ng, Jun-Ping / He, Xiangnang (2013): „Mining Scientific Terms and their Definitions: A Study of the ACL Anthology". EMNLP 2013, 18.-21. Oktober 2013, Seattle, USA, 780-790, http://aclweb.org/anthology/D/D13/D13-1073.pdf, abgerufen am 25.08.2014.

Jongejan, Bart / Dalianis, Hercules (2009): „Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike". ACL-IJCNLP 2009, 2.-7. August 2009, Suntec, Singapur, 145-153, http://www.aclweb.org/anthology/P09-1017, abgerufen am 25.08.2014.

Kalgren, Jussi (2010): „Conventions and Mutual Expectations. Understanding Sources for Web Genres". Mehler, Alexander / Sharoff, Serge / Santini, Marina (Hg.): *Genres on the Web. Computational Models and Empirical Studies.* (Text, Speech and Language Technology 42). Dordrecht: Springer, 33-46.

Keller, Nicole (2006): *Neue Wege in der Hilfsmittelkunde in der Übersetzungswissenschaft. Zur Herleitung webbasierter Terminologiedatenbanken im Kontext von CAT-Systemen*. (Heidelberger Studien zur Übersetzungswissenschaft 5). Trier: WVT.

Kilgarriff, Adam / Grefenstette, Gregory (2003): „Web as Corpus. Introduction to the Special Issue on Web as Corpus". Computational Linguistics 29 (3), 333-347.

Kirschbaum, Ernst-Georg (2001): *Grammatik der russischen Sprache.* Berlin: Volk und Wissen.

Kremer, Gerhard / Abel, Andrea (2010): „Semantic Relations in Cognitive eLexicography". Euralex XIV, 6.-10 Juli 2010, Leeuwarden, Niederlande, 380-388, http://www.eurac.edu/staff/AAbel/Documents/020_1_KREMER%20ABEL_Semantic%20Relations%20in%20Cognitive%20eLexicography_EURALEX_2010.pdf, abgerufen am 25.08.2014.

Kübler, Sandra (2010): „Baumbanken". Carstensen, Kai-Uwe / Ebert, Christian / Ebert, Cornelia / Jekat, Susanne J. / Klabunde, Ralf / Langer, Hagen (Hg.): *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Heidelberg: Spektrum Akademischer Verlag, 492-503.

Kunze, Claudia / Lemnitzer, Lothar (2007): *Computerlexikographie. Eine Einführung*. Tübingen: Narr.

Kustova, Galina I. / Padučeva, Jelena V. (1994): „Slovar' kak leksičeskaja baza dannych" (= Das Wörterbuch als lexikalische Datenbank). Voprosy Jazykoznanija 4, 96-106.

Landis, J. Richard / Koch, Gary G. (1977): „The measurement of observer agreement for categorical data". Biometrics 33(1), 159–174.

Levin, Beth (1993): *English Verb Classes and Alternations. A Preliminary Investigation.* Chicago: University of Chicago Press.

L'Homme, Marie-Claude (2004): *La terminologie: principes et techniques.* Montréal: Les Presses de l'Université de Montréal.

Löbner, Sebastian (2003): *Semantik. Eine Einführung.* Berlin/New York: de Gruyter.

Longworth, Guy (2006): „Definitions: Uses and Varieties of". Brown (Hg.): *Encyclopedia of Language and Linguistics,* 409-412.

Malaisé, Véronique / Zweigenbaum, Pierre / Bachimont, Bruno (2005): „Mining defining contexts to help structuring differential ontologies". Terminology 11 (1), 21-53.

Manning, Christoper D. / Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marshman, Elizabeth (2007): „Towards strategies for processing relationships between multiple relation participants in knowledge patterns: an analysis in English and French". Terminology 13 (1), 1-34.

Marshman, Elizabeth (2008): „Expressions of uncertainty in candidate knowledge-rich contexts". Terminology 14 (1), 124-151.

Mehler, Alexander (2005): „Eigenschaften der textuellen Einheiten und Systeme". Köhler, Reinhard / Altmann, Gabriel / Piotrowski, Rajmund G. (Hg.): *Quantitative Linguistik. Ein internationales Handbuch.* (Handbücher zur Sprach- und Kommunikationswissenschaft 27). Berlin: de Gruyter, 325-348.

Mel'čuk, Igor' A. (1988): *Dependency Syntax: Theory and Practice.* (SUNY series in linguistics). Albany: State University of New York Press.

Mel'čuk, Igor' A. / Clas, André / Polguere, Alain (1995): *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Meyer, Ingrid (2001): „Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework". Bourigault/Jacquemin/L'Homme (Hg.): *Recent Advances in Computational Terminology*, 279-302.

Meyer, Ingrid / Skuce, Douglas / Bowker, Lynne / Eck, Karen (1992): „Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base". COLING 1992, 23.-28. August 1992, Nantes, Frankreich, 956-960, http://aclweb.org/anthology/C/C92/C92-3146.pdf, abgerufen am 25.08.2014.

Mosler, Karl / Schmid, Friedrich (2006[2]): *Wahrscheinlichkeitsrechnung und schließende Statistik.* Berlin: Springer.

Muresan, Smaranda / Klavans, Judith (2002): „A Method for Automatically Building and Evaluating Dictionary Resources". LREC 2002, 29.-31. Mai 2002, Las Palmas, Spanien, 231-234, http://www.lrec-conf.org/proceedings/lrec2002/, abgerufen am 25.08.2014.

Murphy, M. Lynne (2003): *Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms*. Cambridge: Cambridge University Press.

Murphy, Kevin P. (2012): *Machine Learning. A Probabilistic Perspective.* Cambridge, Massachusetts/London: MIT Press.

Navigli, Roberto / Velardi, Paola (2010): „Learning Word-Class Lattices for Definition and Hypernym Extraction". ACL 2010, 11.-16. Juli 2010, Uppsala, Schweden, 1318-1327, http://www.aclweb.org/anthology/P10-1134, abgerufen am 25.08.2014.

Neumann, Stella / Hansen-Schirra, Silvia (2012): „Corpus methodology and design". Hansen-Schirra, Silvia / Neumann, Stella / Steiner, Erich (Hg.): *Cross-Linguistic Corpora For The Study of Translations. Insights from the language pair English-German.* (Text, Translation, Computational Processing 11). Berlin/Boston: de Gruyter, 21-33.

Nistrup Madsen, Bodil / Sandford Pedersen, Bolette / Erdman Thomsen, Hanne (2000): „Defining Semantic Relationships for OntoQuery". Ontology-based interpretation of NP's Workshop, Koldingfjord, Dänemark, 17.-18. Januar 2000, http://www.google.lv/url?sa=t&rct=j&q=defining%20semantic%20relationships%20for%20ontoquery&source=web&cd=1&ved=0CEkQFjAA&url=http%3A%2F%2Fontoquery.dk%2Fpublications%2Fdocs%2FDefining.doc&ei=hCroT9C2DsSh4gT_g9WiAQ&usg=AFQjCNE5iDnib0Qqip7rK2acZovbTX_TPg, abgerufen am 05.05.2014.

Nivre, Joakim / Hall, Johan / Nilsson, J. / Chanev, A. / Eryigit, G. / Kübler, Sandra / Marinov, S. / Marsi, Erwin (2007): „MaltParser: A language-independent system for data-driven dependency parsing". *Natural Language Engineering*, 13(2), 95-135.

Nivre, Joakim / Boguslavsky, Igor M. / Iomdin, Leonid L. (2008): „Parsing the SYNTAGRUS Treebank of Russian". COLING 2008, 18.-22. August 2008, Manchester, England, 641-648, http://www.aclweb.org/anthology/C08-1081.pdf, abgerufen am 25.08.2014.

Nord, Christiane (1995[3]): *Textanalyse und Übersetzen. Theoretische Grundlagen, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse.* Heidelberg: Groos.

Nuopponen, Anita (1994): „Wüster revisited: On Causal Concept Relationships and Causal Concept Systems". 9th European Symposium on LSP, Bergen. Bd. 2, Norwegen, 2.-6. August 1993, 532-539, ttp://lipas.uwasa.fi/~atn/papers/artikkelit/LinkedDocuments/Nuopponen_Causal_LSP94.pdf, abgerufen am 25.08.2014.

Oberholzer, Mirjam (2002): *Terminologische Definitionen: Form, Funktion, Extraktion.* Lizenziatsarbeit. Zürich: Philosophische Fakultät.

Pamjatka (2011): *SinTagRus: Sintaksičeski razmečennyj korpus russkich tekstov. Informacija dlja pol'zovatelja* (= SynTagRus: Ein syntaktisch annotiertes Korpus russischer Texte). Im Korpus beiliegendes Informationsblatt).

Pearson, Jennifer (1998): *Terms in Context*. (Studies in Corpus Linguistics 1). Amsterdam/Philadelphia: John Benjamins.

Pribbenow, Simone (2002): „Meronymic relationships: From Merology to Complex Part-Whole Relationships". Green / Bean / Myaeng (Hg.): *The Semantics of Relationships*, 35-50.

Picht, Heribert (1987): „Terms and their LSP Environment – LSP Phraseology". Meta 32 (2), 149-155.

Picht, Heribert (1997): „Erarbeitung und Anwendung begrifflicher Strukturierungen". Hermes 18, 33-50.

Przepiórkowski, Adam / Degórski, Łukasz / Spousta, Miroslav / Simov, Kiril / Osenova, Petya / Lemnitzer, Lothar / Kuboň, Vladislav / Wójtowicz, Beata (2007): „Towards the Automatic Extraction of Definitions in Slavic". BSNLP Workshop, ACL 2007, 29.6.2007, Prag, Tschechische Republik, 43-50, http://langtech.jrc.it/BSNLP2007/m/BSNLP-2007-proceedings.pdf, abgerufen am 25.08.2014.

Pustejovsky, James (1991): „The generative lexicon". Computational Linguistics 17 (4), 209-441.

Quasthoff, Uwe / Richter, Matthias/ Biemann, Christian (2006): „Corpus Portal for Search in Monolingual Corpora". LREC 2006, 22.-28. Mai 2006, Genua, Italien, 1799–1802, http://www.lrec-conf.org/proceedings/lrec2006/, abgerufen am 25.08.2014.

R Development Core Team (2012): *R. A language and environment for statistical computing.* Wien: R foundation for statistical computing, http://web.mit.edu/r_v3.0.1/fullrefman.pdf, abgerufen am 25.08.2014.

Rafferty, Anna N./ Manning, Christopher D. (2008): „Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines". Workshop on Parsing German, ACL 2008, 20. Juni 2008, Columbus, USA, 40-46, http://www.aclweb.org/anthology/W08-1006.pdf, abgerufen am 25.08.2014.

Reichenberger, Klaus (2010): *Kompendium Semantische Netze. Konzepte, Technologie, Modellierung*. Berlin/Heidelberg: Springer.

Reiplinger, Melanie / Schäfer, Ulrich / Wolska, Magdalena (2012): „Extracting glossary sentences from scolarly articles: A comparative analysis of pattern bootstrapping and deep analysis". Workshop on Rediscovering 50 Years of Discoveries, ACL 2012, 10. Juli 2010, Jeju, Korea, 55-65, http://www.aclweb.org/anthology/W12-3206, abgerufen am 25.08.2014.

Rirdance, Signe / Vasiļjevs, Andrejs (2006): *Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project*. Riga: Tilde.

Roelcke, Thorsten (2005[2]): *Fachsprachen.* Berlin: Erich Schmidt.

Roelcke, Thorsten (2013): *Definitionen und Termini. Quantitative Studien zur Konstituierung von Fachwortschatz.* (Linguistik – Impulse und Tendenzen 48). Berlin/Boston: de Gruyter.

Roche, Christophe (2012): „Should *Terminology Principles* be re-examined?". TKE 2012, Madrid, Spanien, 19.-22. Juni 2012, 17-32, http://oeg-lia3.dia.fi.upm.es/c/document_library/get_file?uuid=9a9897a4-9382-4b4c-90dd-105829b09c84&groupId=10157, abgerufen am 25.08.2014.

Rosenthal, Robert / Rosnow, Ralph L. / Rubin, Donald B. (2000): *Contrasts and Effect Sizes in Behavioral Research. A Correlational Approach.* Cambridge: Cambridge University Press.

Sachs, Lothar / Hedderich, Jürgen (2009[13]): *Angewandte Statistik. Methodensamlung mit R.* Berlin: Springer.

Schmid, Helmut (1994): „Probabilistic Part-of-Speech Tagging Using Decision Trees". International Conference on New Methods in Language Processing, Manchester, England, 44–49; http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf; abgerufen am 25.08.2014.

Schulte im Walde, Sabine /Brew, Chris (2002): „Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information". ACL 2002, 6.-12. Juli 2002, Philadelphia, USA, 223-230, http://www.aclweb.org/anthology/P/P02/P02-1029.pdf, abgerufen am 25.08.2014.

Schumacher, Helmut (Hg.) (1986): *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben.* (Schriften des Instituts für Deutsche Sprache 1). Berlin/New York: de Gruyter.

Séaghdha, Diarmuid Ó / Copestake, Anne (2009): „Using lexical and relational similarity to classify semantic relations". EACL 2009, 30. März – 3. April 2009, Athen, Griechenland, 621-629, http://www.aclweb.org/anthology/E09-1071, abgerufen am 25.08.2014.

Sebeok, Thomas Albert (1986): *Encyclopedic Dictionary of Semiotics*. *Bd. 1*. (Approaches to Semiotics 73). Berlin/New York: de Gruyter.

Seeker, Wolfgang / Kuhn, Jonas (2012): „Making Ellipses Explicit in Dependency Conversion for a German Treebank". LREC 2012, 21.-27. Mai 2012, Istanbul, Türkei, 3132–3139, http://www.lrec-conf.org/proceedings/lrec2012/index.html, abgerufen am 25.08.2014.

Sharoff, Serge (2006): „Creating General-Purpose Corpora Using Automated Search Engine Queries". Baroni, Marco / Bernardini, Silvia (Hg.): *WaCky! Working papers on the Web as Corpus*. (Studi Interdisciplinari su Traduzione, Lingue e Culture). Bologna: Gedit, 63-98, http://www.lrec-conf.org/proceedings/lrec2008/index.html, abgerufen am 25.08.2014.

Sharoff, Serge / Kopotev, Mikhail / Erjavec, Tomaž/ Feldmann, Anna / Divjak, Dagmar (2008): „Designing and evaluating a Russian tagset". LREC 2008, 26. Mai-1. Juni 2008, Marrakesch, Marokko, 279-285.

Sharoff, Serge / Nivre, Joakim (2011): „The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge." Dialogue 2011, 25.-29. Mai 2011, Bekasovo, Russland, 591-604.

Sierra, Gerardo / Alarcón, Rodrigo / Aguilar, César / Bach, Carme (2008): „Definitional verbal patterns for semantic relation extraction". Terminology 14 (1), 74-98.

Sinclair, John (1991): *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Skut, Wojciech / Brants, Thorsten / Krenn, Brigitte / Uszkoreit, Hans (1997): „Annotating Unrestricted German Text". *Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft,* 8.-10. Oktober 1997, Heidelberg, Deutschland, http://www.coli.uni-saarland.de/~thorsten/publications/Skut-ea-DGfS97.pdf, abgerufen am 25.08.2014.

Sowa, John (1984): *Conceptual Structures: Information Processing in Mind and Machine*. Reading: Addison-Wesley.

Stanaitytė, Greta (2005): *Alltagsdefinitionen und ihre Funktionen.* Dissertation. Mannheim: Universität Mannheim.

Storrer, Angelika / Wellinghoff, Sandra (2006): „Automated detection and annotation of term definitions in German text corpora“. LREC 2006, 22.-28. Mai 2006, Genua, Italien, 2373-2376, http://www.lrec-conf.org/proceedings/lrec2006/, abgerufen am 25.08.2014.

Švedova, Natalija Jul'evna (2007): *Russkij semantičeskij slovar'. Vol 4: Glagol* (= Russisches semantisches Wörterbuch. Band 4: Verb). Moskva: Azbukovnik.

Teich, Elke (2003): *Cross-Linguistic Variation in System and Text.* (Text, Translation, Computational Processing 5). Berlin/New York: de Gruyter.

Teich, Elke / Fankhauser, Peter (2010): „Exploring a corpus of scientific texts using data mining". Gries, Stefan / Wulff, Stefanie / Davies, Mark (Hg.): *Corpus-linguistic applications. Current studies, new directions.* (Language and Computers – Studies in Practical Linguistics 71). Amsterdam/New York: Rodopi, 233-247.

Temmerman, Rita (2000): *Towards New Ways of Terminology Description. The sociocognitive approach*. Amsterdam/Philadelphia: John Benjamins.

Teubert, Wolfgang / Čermáková, Anna (2007): *Corpus Linguistics: A Short Introduction*. London: Continuum.

Tuldava, Juhan (2005): „Stylistics, author identification". Köhler/ Altmann / Piotrowski (Hg.): *Quantitative Linguistik,* 368-387.

Van Campenhoudt, Marc (2001): „Pour une Approche Sémantique du Terme et de ses Équivalents". International Journal of Lexicography 14 (3), 181-209.

Vossen, Piek /Meijs, Willem / den Broeder, Marianne (1989): „Meaning and structure in dictionary definitions“. Boguraev / Briscoe (Hg.): *Computational Lexicography for Natural Language Processing,* 171-192.

Walter, Stephan (2010): *Definitionsextraktion aus Urteilstexten.* Dissertation. Saarbrücken: Universität des Saarlandes.

Walter, Stephan / Pinkal, Manfred (2006): „Automatic Extraction of Definitions from German Court Decisions". Workshop on Information Extraction beyond the Document, COLING/ACL 2006, 22. Juli 2006, Sydney, Australien, 20-28, http://www.aclweb.org/anthology/W/W06/W06-0203.pdf, abgerufen am 25.08.2014.

Waterman, Scott A. F. (1998): „Distinguished Usage". Fellbaum, Christiane (Hg.): *WordNet: An electronic lexical database*. Cambridge, Massachusetts/London: MIT Press, 143-172.

Weber, Nico (1999): *Die Semantik von Bedeutungsexplikationen.* (Sprache, Sprechen und Computer 3). Frankfurt/Main: Lang.

Westerhout, Eline (2009): „Definition Extraction Using Linguistic and Structural Features“. First Workshop on Definition Extraction, RANLP 2009, 18.9.2009, Borovets, Bulgarien, 61-67, http://www.aclweb.org/anthology/W09-4410, abgerufen am 25.08.2014.

Westerhout, Elizabeth Nicoline (2010): *Definition extraction for glossary creation. A study on extracting definitions for semi-automatic glossary creation in Dutch.* Dissertation. Utrecht: Universität Utrecht.

Wiegand, Herbert Ernst (1989): „Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch“. Hausmann, Franz Josef (Hg.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie.* (Handbücher zur Sprach- und Kommunikationswissenschaft 5). Bd. 1. Berlin/New York: de Gruyter, 530-588.

Wimmer, Gejza (2005): „The type-token relation". Köhler/ Altmann / Piotrowski (Hg.): *Quantitative Linguistik,* 361-368.

Witten, Ian H. / Frank, Eibe / Hall, Mark A. (2011[3]): *Data Mining. Practical Machine Learning Tools and Techniques.* Amsterdam: Elsevier.

Wolf, Christof / Best, Henning (2010): „Logistische Regression". Wolf, Christof/Best Henning (Hg.): *Handbuch der sozialwissenschaftlichen Datenanalyse.* Wiesbaden: VS Verlag für Sozialwissenschaften, 827-854.

Wollschläger, Daniel (2012[2]): *Grundlagen der Datenanalyse mit R. Eine anwendungsorientierte Einführung.* Berlin/Heidelberg: Springer.

Wright, Sue Ellen (2007): „Coping with indeterminancy. Terminology and knowledge representation resources in digital environments". Antia, Bassey Edem (Hg.): *Indeterminacy in Terminology and LSP. Studies in honour of Heribert Picht*. (Terminology and Lexicography Research and Practice 8). Amsterdam/Philadelphia: John Benjamins, 157-179.

Wüster, Eugen (1991[3]): *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* (Abhandlungen zur Sprache und Literatur 20). Bonn: Romanistischer Verlag.

Xiao, Richard (2008): „Well-known and influential corpora". Lüdeling, Anke / Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. *Bd. 1*. Berlin/New York: de Gruyter, 383-457.

Xu, Fei-Yu (2007): *Bootstrapping Relation Extraction from Semantic Seeds*. Dissertation. Saarbrücken: Universität des Saarlandes.

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno (1997): *Grammatik der deutschen Sprache.* Bd. 1-3. (Schriften des Instituts für Deutsche Sprache 7.1-7.3). Berlin: de Gruyter.

Zinsmeister, Heike / Kuhn, Jonas / Schrader, Bettina / Dipper, Stefanie (2001): *TIGER Transfer. From LFG Structures to the TIGER Treebank.* Technischer Bericht. Stuttgart: Universität Stuttgart, http://www.researchgate.net/publication/2376250_Tiger_Transfer_-_From_LFG_Structures_to_the_TIGER_Treebank, abgerufen am 25.08.2014.

Zolotova, G. A. / Onipenko, N. K. / Sidorova, M. Ju. (1998): *Kommunikativnaja grammatika russkogo jazyka* (= Kommunikative Grammatik der russischen Sprache). Moskva: RAN.

Anhang 1: Unterlagen zur Annotation wissenshaltiger Kontexte für den Goldstandard

Forschungshintergrund

- Was sind wissenshaltige Kontexte?
- Wozu braucht man sie?
- Wie extrahiert man sie?

Forschungshintergrund

- Terminographie benötigt:
  - Linguistische Information („Wie heißt XY in Sprache b?“)
    -> Erwerb sprachlichen Wissens (Wort nachschlagen)
  - Semantische Information („Was ist XY eigentlich?“)
    -> Erwerb von Fachwissen (Fachgebiet verstehen und richtig übersetzen)

Forschungshintergrund

- Wissenshaltige Kontexte liefern *semantische Informationen* über *Begriffe:*
  - natürlich vorkommende Textbestandteile (etwa aus einem Korpus von Web-Texten)
  - Helfen dem Leser, einen fachspezifischen Begriff zu verstehen, und zwar:
    - dessen intrinsische Eigenschaften (Attribute) – Größe, Verhalten ...
    - in welcher Beziehung er zu anderen Begriffen steht (welche semantischen Relationen er mit Nachbarbegriffen eingeht)
  - Information muss explizit gegeben sein
  - zeitliche Gültigkeit der Information muss klar erkennbar sein

Forschungshintergrund

- Welche semantischen Beziehungen beschreiben Kontexte?
  - Oberbegriff - Unterbegriff
  - Teil – Ganzes
  - Prozesse
  - Positionen
  - Ursache – Wirkung
  - Begriff – Herkunft
  - Begriff – Funktion
  - Synonyme, Umschreibungen ...

Forschungshintergrund

- Beispiele:

Oberbegriff: Die Mantelturbine gehört zu den konzentrierenden Systemen.

Teile: Ein Wechselstromrichter besteht aus elektronischen Komponenten (hauptsächlich aus IGBTs -Thyristoren und Dioden) ...

Prozesse: Das Getriebe 2, das von der Rotorwelle angetrieben wird, ...

Forschungshintergrund

- Beispiele:

Position: Im Läufer befindet sich eine Wicklung, die direkt oder über einen Widerstand kurzgeschlossen ist.

Kausalität: Eine Verstärkung des Auftriebes wird durch die Erhöhung des so genannten „Anstellwinkels“ bewirkt, ...

Forschungshintergrund

- Beispiele:

Funktion: Die auch bei Windkraftanlagen mit mehr als 100 Metern Höhe vorgeschriebene Hindernisbefeuerung dient der Sicherheit des Flugverkehrs.

Restkategorie: Onshore bedeutet, dass sich die Windkraftanlage an Land befindet, offshore bezeichnet Anlagen, die sich auf dem Meer befinden.

Forschungshintergrund

-> Validitätskriterien:

- Der Kontext kann einer der relevanten semantischen Relationen zugeordnet werden.
- Der Kontext kann einem Zielterminus zugeordnet werden.
- Die zeitliche Gültigkeit der durch den Kontext vermittelten Information ist nachvollziehbar.
- Der Kontext kann einem Fachgebiet zugeordnet werden.
- Die durch den Kontext vermittelte Information ist generalisierbar oder trägt einen relevanten Teilaspekt zum Verständnis des Begriffs bei.
- Der Kontext enthält relevante Informationen zu Attributen oder semantischen Relationen.

Anhang 2: Fehlannotation in den deutschen Korpora

a) Erstannotation Sub-EKD

	Satz	Erläuterung
Fehlannotation von Annotator De1	Ein Windkraftanlagenblatt besteht aus zwei Halbschalen (der Saug- und der Druckseite), welche durch mehrere, interne Stege verbunden sind Die Halbschalen werden mit Glas- oder Kohlefaser mit mehreren Lagen belegt und in Polyester und Epoxydharz getränkt.	Aufgrund des fehlenden Satzzeichens weist der Kontext-Kandidat eine sprachliche Unregelmäßigkeit auf, entspricht sonst aber den erarbeiteten Validitätskriterien, wobei allerdings nur der erste Teilsatz in den Goldstandard aufgenommen wird.
	Die Energierücklaufzeit beschreibt die Zeit, die vergeht, bis ein Kraftwerk genauso viel Energie erzeugt hat, wie zu seiner Produktion, Transport, Errichtung, Betrieb usw. benötigt wurde.	Der Satz entspricht den formalen Kriterien einer Definition.
	Konzentrierte Luftwirbel treten als sogenannte Randwirbel bei der Umströmung eines Tragflügels auf.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
	Das Tension-Leg-Konzept beschreibt eine schwimmende Gründdungsstruktur, die tauchend über Zugabspannungen im Meeresboden verankert wird.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
Fehlannotation von Annotator De2	Bei großen Windkrafträdern spricht man hierbei vom Maschinenhaus, das im Fachjargon auch "Gondel" genannt wird.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Die zylindrische Einleitflächenkonstruktion 11 besteht aus einem zylindrischen Einleitflächenaußenring 14, welcher eine unterschiedliche Anzahl von Etagen aufweisen kann.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Die zylindrische Einleitflächenkonstruktion 11 sitzt auf dem Dach 18 des Maschinenraumes 8 auf.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Die zylindrische Einleitflächenkonstruktion 11 sitzt auf dem Maschinenraum 8.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

	Mehr als 90% der weltweit betriebenen Windkraftanlagen sind mit direkt netzgekoppelten Asynchrongeneratoren ausgerüstet.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Meist sind größere Windkonverter mit einer automatischen Rotorblattverstellung ausgerüstet.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Theoretische und praktische Nutzungspotenziale Das technisch erschließbare einheimische Potenzial der energetischen Nutzung wird unterschieden in Onshore - Nutzung und Offshore - Nutzung.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Um die erheblich stärkeren Winde auf See nutzen zu können, wird in Deutschland vermehrt die Errichtung von Windparks auf dem offenen Meer, sogenannte Offshore-Windparks, geplant.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Zur Stromerzeugung werden heute nur noch Synchrongeneratoren verwendet (und kaum noch Asynchrongeneratoren).	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

b) Erstannotation Sub-TKD

	Satz	**Erläuterung**
Fehlannotation von Annotator De1	Das Magische Auge ist eine Spezialröhre , die ein elektrisches Steuersignal in ein Leuchtsignal umsetzt , dessen sichtbare Ausdehnung von der angelegten Steuersignalspannung abhängt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen erfüllt der Kontext-Kandidat die formalen Kriterien für eine Definition.
	Das Verhalten von Synchronmotoren am Netz wird oft statt durch die Regelkennlinien durch die wegen ihrer Form sogenannten V-Kurven beschrieben.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.

	Der Tripelpunkt des Wasserstoffs , bei dem seine drei Aggregatzustände gleichzeitig vorkommen , ist einer der Fixpunkte der Internationalen Temperaturskala .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
Fehlannota-tion von Annotator De2	Das Ziel der europäischen Normung ist die Harmonisierung der nationalen Normen in den Mitgliedsländern durch einheitliche Einführung von Europäischen Normen .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
	Das Verhältnis von tatsächlicher zu Carnot - Leistungszahl ist der Gütegrad η = ε / ε C .	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Der Vierpolparameter Durchgriff beschreibt die Rückwirkung eines sich ändernden Anodenpotenzials auf den Anodenstrom .	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Die Netzstrom-Erregerstrom-Kennlinien bei verschiedenen Leistungsfaktoren nennt man die Regelkennlinien der Synchronmaschine.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

	Die Trennlinie zwischen fester und flüssiger Phase , also die Kurve der Schmelzpunkte , nennt man Schmelzkurve , die Trennlinie zwischen Flüssigkeit und Gas nennt man Siedepunktskurve .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
	Die verschiedenen Röhrentypen werden durch ihre Funktion und die Anzahl und Anordnung der Elektroden unterschieden , außerdem durch unterschiedliche Stromversorgung (Netzröhren , Batterieröhren und Niederspannungsröhren) .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Die zu lösende objektive technische Aufgabe besteht darin , über die Änderung oder Anpassung des nächstliegenden Stands der Technik , die technischen Wirkungen zu erzielen , welche die Erfindung über den Stand der Technik mit sich bringt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.

	Direkt geheizte Kathoden werden noch heute bei Senderöhren , Gleichrichterröhren und Magnetrons eingesetzt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Erdgas wird heute zur Stromerzeugung hauptsächlich in Gasturbinen - oder GuD - Kraftwerken (Gas - und Dampfkraftwerken) eingesetzt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Erst ab der so genannten Schmelztemperatur wird die Schwingungsamplitude der Teilchen so groß , dass die Gitterstruktur teilweise zusammenbricht .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

	Fluoreszenzanzeigen werden als Anzeigeeinheiten bei Heimelektronik - Geräten wie Videorekordern oder DVD - Playern nach wie vor häufig verwendet .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Heute finden sich Trioden vor allem in Vorverstärkerstufen von Hi - End - Audiogeräten – in Endstufen werden sie nur für höchste Frequenzen , ansonsten aufgrund des geringeren Verstärkungsfaktors nur bei besonderen Anforderungen an die Linearität (Kennlinie ist Kurve erster Ordnung) eingesetzt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Im Allgemeinen besteht ein Wicklungsstrang aus einer Spule, die mehr als die hier gezeichneten 2 Windungen besitzt.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Indirekt geheizte Röhren werden zur Verstärkung kleiner Signale (ältere Fernseher , Messgeräte und Radioempfänger) und heute noch in Audioverstärkern eingesetzt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

	Wasserstoff bildet mit Kohlenstoff die kovalenten Kohlenwasserstoffe , deren Studium sich die Kohlenwasserstoffchemie verschrieben hat .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Wie Sie wissen, haben Dreiphasen-Induktionsmotoren drei Wicklungsstränge im Ständer, die von drei zeitlich gegeneinander verschobenen Strömen durchflossen werden.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Zur Beobachtung dieser Bedingungen dienen besondere Messgeräte, sogenannte Synchronoskope, die melden, wann die Bedingungen für ein Synchronisieren erfüllt sind, bzw. anzeigen, wie Erregung und Drehzahl des Generators geändert werden müssen, damit die Bedingungen erfüllt werden.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
Fehlannota-tion beider Annotatoren	Der nächstliegende Stand der Technik ist eine einzige Offenbarung , die den erfolgversprechendsten Ausgangspunkt für eine naheliegende Entwicklung darstellt , die zur beanspruchten Erfindung führt .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen erfüllt der Kontext-Kandidat die formalen Kriterien für eine Definition.

	Mündliche Beschreibungen und im Ausland erfolgte Benutzungshandlungen zählen somit nicht zum Stand der Technik .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen erfüllt der Kontext-Kandidat die formalen Kriterien für eine Definition.

c) Zweitannoation Sub-EKD

	Satz	**Erläuterung**
Fehlannotation von Annotator De3	Die Windrichtungsnachführung wird durch hydraulische Motoren oder Elektromotoren (Azimutantrieb) gewährleistet.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
Fehlannotation von Annotator De4	Bei voll gedämpften Rotorsystemen befindet sich an jedem Rotorblattgriff je ein Gelenk, welches dem Rotor ermöglicht, sich nach oben und unten bzw. nach vorne und hinten "frei" zu bewegen, ohne diese Bewegung über den Rotorkopf auf den Rumpf zu übertragen.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
	Das Bucket-Fundament wird als eimerartiges Stahlfundament ausgeführt, das mit der Öffnung nach unten durch ein Vakuumverfahren in den Meeresboden eingebracht wird.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen erklärenden Kontext.
	Die wesentlichsten Einflußgrößen für die Dimensionierung des Fundaments, mit dem der Turm und damit die Windkraftanlage im Untergrund verankert wird, sind die Anlagengröße, die meteorologischen und betrieblichen Belastungen und die örtlichen Bodenverhältnisse.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Infraschall, also Schall mit Frequenzen zwischen 0 Hz und 16 Hz, tritt z.B. im Zusammenhang mit Erdbeben oder tektonischen Verschiebungen als seismische Welle auf.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

d) Zweitannotation Sub-TKD

	Satz	Erläuterung
Fehlannotation von Annotator De3	Bildverstärker (Restlichtverstärker) und Bildaufnahmeröhren sind Bildwandlerröhren , welche vor allem zur Aufnahme und Verstärkung von Licht dienen .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen erfüllt der Kontext-Kandidat die formalen Kriterien für eine Definition.
	Kernzerfallsgeneratoren wurden früher u . a . für Herzschrittmacher verwendet , heutzutage werden sie vornehmlich als Stromquelle und Heizung für Raumsonden im äußeren Sonnensystem eingesetzt , da dort Solarzellen keine ausreichende Leistung und keine Wärme liefern .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
Fehlannotation von Annotator De4	Ammoniak ist aus einzelnen Molekülen aufgebaut , die jeweils aus einem Stickstoff - und drei Wasserstoffatomen bestehen .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

	Das heißt , es fehlt an Erfindungshöhe , wenn man von einem Fachmann erwarten kann , dass er , ausgehend vom Stand der Technik , auf diese Lösung alsbald und mit einem zumutbaren Aufwand gekommen wäre , ohne erfinderisch tätig zu werden .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Die vereinheitlichten europäischen Patentgesetze bestimmen – neben anderem – einhellig : Eine Erfindung ist nur dann patentfähig , „ wenn sie sich für den Fachmann nicht in naheliegender Weise aus dem Stand der Technik ergibt “ (§ 4 Satz 1 , Art . 56 des Europäischen Patentübereinkommens = EPÜ) .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Nur wenige Mikroorganismen sind in der Lage , Ammoniak in der sogenannten Stickstofffixierung direkt aus dem Stickstoff der Luft zu gewinnen .	Von kleinen, wahrscheinlich auf den Extraktionsprozess aus dem Internet rückführbaren Unregelmäßigkeiten in Bezug auf die Zeichensetzung abgesehen entspricht der Kontext-Kandidat den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

Anhang 3: Fehlannotation in den russischen Korpora

a) Erstannotation Sub-EKR

	Satz	Erläuterung
Fehlannotation von Annotator Ru1	Воздушный фильтр предназначен для очистки воздуха, посту-пающего от компрессора в пневматическую систему от влаги и от масла.	Wahrscheinlich aufgrund von Konvertierungsfehlern weist der Satz eine Unregelmäßigkeit in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.
	Манометр позволяет проверять давление воздуха, как в воздуш-ных баллонах, так и в тормозных камерах системы пневматического привода.	Wahrscheinlich aufgrund von Konvertierungsfehlern weist der Satz eine Unregelmäßigkeit in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.
	Один контур гидропривода обеспечивает работу правого переднего и левого заднего тормоз-ных механизмов, другого – левого переднего и правого заднего.	Wahrscheinlich aufgrund von Konvertierungsfehlern weist der Satz zwei Unregelmäßigkeiten in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.
	Привод стояночного тормоза состоит из рычага, регули-ровочной тяги, уравнителя, троса, рычага ручного привода колодок и разжимной план-ки.	Wahrscheinlich aufgrund von Konvertierungsfehlern weist der Satz zwei Unregelmäßigkeiten in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.

	Регулятор давления автоматически поддерживает необходимое давле-ние сжатого воздуха в системе.	Wahrscheinlich aufgrund von Konvertierungs-fehlern weist der Satz zwei Unregel-mäßigkeiten in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.
	Регулятор давления служит для регулирования давления в гидроприводе тормоз-ных механизмов задних колёс в зависимости от нагрузки на заднюю ось автомобиля.	Wahrscheinlich aufgrund von Konvertierungs-fehlern weist der Satz zwei Unregel-mäßigkeiten in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.
	Современный автомобиль оборудуется рабочей, запасной, стояночной и вспомо-гательной тормозными системами.	Wahrscheinlich aufgrund von Konvertierungs-fehlern weist der Satz zwei Unregel-mäßigkeiten in der sprachlichen Oberfläche auf, entspricht aber dennoch den Kriterien für einen assoziativen Kontext.
Fehlannota-tion beider Annotatoren	Вспомогательная ТС начинает действовать при разгерметизации одного из рабочих контуров (вытекает тормозная жидкость).	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Катушка зажигания вызывает искровой разряд (вспышку) из свечей зажигания на каждом рабочем такте (для цилиндра на такте сжатия и для цилиндра на такте выпуска отработавших газов).	Der Satz entspricht trotz inhaltlicher Mängel den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

b) Zweitannotation Sub-EKR

	Satz	**Erläuterung**
Fehlannota-tion von Annotator Ru4	Колесные тормоза колодочного типа с гидравлическим управлением (рис. 26) устанавливают на легких и средних автогрейдерах.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

c) Zweitannotation Sub-TKR

	Satz	**Erläuterung**
Fehlannotation von Annotator Ru3	Сталь содержит также вредные примеси (фосфор, серу, газы - несвязанный азот и кислород).	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Например, к сварным конструкциям относятся кузова автомобилей, фюзеляжи самолётов и т.д.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Все параметры режима сварки можно разделить на основные и дополнительные.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
Fehlannotation von Annotator Ru4	Солнечная энергия может быть использована для теплоснабжения (горячего водоснабжения, отопления), сушки различных продуктов и материалов, в сельском хозяйстве, в технологических процессах в промышленности.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.
	Тепловое загрязнение влияет на репродуктивные способности организмов, на их способность добывать пищу, сопротивляться болезням и избегать хищников.	Der Satz entspricht den in Kapitel 2 dargelegten Kriterien für einen assoziativen Kontext.

Anhang 4: Falsche Positive in den deutschen Korpora

a) Sub-EKD

Kontext-Kandidat	Fehlerkategorie
Als Vorzugsvariante eines optimal wirkenden vertikalen Einleitflächenelmentes dient die gekrümmte Form mit aerodynamisch gewölbter Kopfform, welche jedoch in keiner Weise der gekrümmten Nasenform der Rotorflügel entsprechen darf.	Semantische Relation unklar
Anlagen mit Heidelberg-Rotoren verfügen über einen neu entwickelten Generator, den so genannten Wanderfeldgenerator.	Information zu allgemein
Auch Asynchromgeneratoren besitzen außen einen feststehenden Stator und einen drehbaren Läufer.	Terminus unklar
Beide Generatoren bestehen außen aus einem Stator und innen aus einen Läufer (auch Anker genannt), der sich auf einer drehbaren Welle befindet.	Verallgemeinerbar-keit fraglich
Bis vor wenigen Jahren wurden große Synchrongeneratoren nur in Verbindung mit Übersetzungsgetrieben verwendet.	Zeitliche Gültigkeit unklar
Das Getriebe 2, das von der Rotorwelle angetrieben wird, die Hydraulikpumpe 3 und der Drehstromgenerator 7 bilden eine zusammengeflanschte Einheit, die auf dem schwimmfähigen Tragwerk befestigt ist.	Semantische Relation unklar
Das Stellglied ist ein Komparator, der die Stellgröße mit einem Dreiecksignal vergleicht und so die Schaltfolgen für die IGBTs erzeugt (Analoge PWM-Erzeugung).	Verallgemeinerbar-keit fraglich
Der Einleitflächenaußenring besitzt eine Ausladung, welche einem Drittel des Rotordurchmessers entspricht.	Verallgemeinerbar-keit fraglich
Der Generator bringt ein Gegenmoment zum Rotor auf.	Semantische Relation unklar
Der gesamte Kopf einer Windkraftanlage ist drehbar angeordnet, um sich immer genau in Windrichtung stellen zu können.	Semantische Relation unklar
Der Hauptgrund hierfür liegt zum einen darin, dass die tragende Struktur einer Windenergieanlage, bestehend aus Turm und Gondel sehr elastisch ist und in Kombination mit dem Rotor ein extrem schwingungsanfälliges System darstellt.	Terminus unklar
Der Kollektivhebel (Pitch) steuert die Taumelscheibe über Hydraulik (Steuerstangen) an und verschiebt diese parallel nach oben oder unten.	Verallgemeinerbar-keit fraglich
Der Planetenträger (auch Steg genannt) (gelb), kann fixiert oder beweglich sein.	Fachlicher Bezug unklar
Der Rotor ist stallgeregelt (dass heißt durch Strömungsabriß) und zusätzlich durch hydraulische Drosselung, die der Hydraulikpumpe 3 nachgeschaltet ist.	Information zu allgemein
Der Segelzug wirkt somit als Bindeglied, damit die Trägheit im Rotor überwunden wird und ein fließender ungestörter Bewegungablauf der Rotation gewährleistet wird.	Semantische Relation unklar

Der so genannte "Diskoeffekt" (Lichtreflexe an den Rotorblättern) spielt ebenfalls keine Rolle mehr, denn schon lange werden auf den Rotorflächen matte, nichtreflektierende Farben aufgetragen.	Semantische Relation unklar
Der Stall-Effekt stellt eine Möglichkeit zur Begrenzung der Leistungsaufnahme aus dem Wind dar.	Information zu allgemein
Der Steuerknüppel (Stick) steuert die Taumelscheibe über Hydraulik (Steuerstangen) an und versetzt diese in eine Schräglage.	Fachlicher Bezug unklar
Der Tornado-Typ, der zu den "Wirbeltürmen" zählt, könnte mit einem Durchmesser von 180 m und 550 m Höhe bis zu 1.000 MW elektrische Energie bei 13,3 m/s Windgeschwindigkeit erzeugen.	Information zu allgemein
Der Unterwasserrotor kann mechanisch, elektrisch, hydraulisch oder mittels einer Kombination dieser Antriebsmöglichkeiten angetrieben werde.	Verallgemeinerbar-keit fraglich
Der Vorteil der Leeläufer wäre die Möglichkeit eine passive Windrichtungsnachführung einzurichten (der Wind lässt die Windkraftanlage sich selbst drehen), um den Preis und Gewicht des Windrichtungsnachführungssystems zu sparen.	Information zu allgemein
Die an sich bekannte Variante der Energietanne hat durch die im unteren Bereich größere Ausladung auch die Funktion eines Vorverdichters.	Fachlicher Bezug unklar
Die Asynchronmaschine ist am Ständer direkt mit dem Stromnetz verbunden.	Verallgemeinerbar-keit fraglich
Die erfindungsgemäße Windkraftanlage ist dazu mit einem idealen Durchströmrotor ausgestattet, welcher drei besonders gestaltete Flügel besitzt	Verallgemeinerbar-keit fraglich
Die Flügel 1, 2 und 3 besitzen an ihrer Frontseite eine gekrümmte Flügelnase 15.	Verallgemeinerbar-keit fraglich
Die Flügelausbildung des Rotors entspricht in der Krümmung der des Schnelläufers nach dem Prinzip Schulz als dreiflügliger Durchströmrotor, welcher aus dem System des Savonius-Rotors hervorging.	Information zu allgemein
Die Frequenzkonstanz wird durch die Flügelgruppenversetzung in den Etagen und durch den Schwungradeffekt erreicht.	Fachlicher Bezug unklar
Die gelenkige Einspannung wird als "schlagend" bezeichnet.	Fachlicher Bezug unklar
Die grundlegendste Kraft ist die Schwerkraft, die jeden Körper auf den Erdmittelpunkt (Erdboden) hin zieht.	Information zu allgemein
Die Landung kann auch mit 0 Knoten durchgeführt werden und wird dann als vertikale Autorotation bezeichnet.	Fachlicher Bezug unklar
Die Musgrove-Anlagen verfügen teilweise über Rotoren mit variabler Geometrie.	Information zu allgemein
Die Rotorflächen, die Flächen der Flügel 1, 2, 3 und die Flächen der Einleitelemente 12, 13 sind starr und unbeweglich ausgeführt, weshalb keinerlei flexible Materialien Einsatz finden.	Verallgemeinerbar-keit fraglich

Die Schallimmissionsprognose hat die Funktion, schon vor Errichtung einer Windenergieanlage anhand der konkreten Gegebenheiten der Örtlichkeit und die technischen Spezifikation der geplanten Anlage eine zuverlässige Aussage darüber zu erlauben, ob die Nachbarn am vorgesehenen Standort Lärmimmissionen ausgesetzt sein werden, die über das von ihnen hinzunehmende Maß hinausgehen.	Dopplung
Die sich drehenden Rotorblätter bewirken ein Flackern des Schattens, den sogenannten Discoeffekt.	Fehlerhafte Information
Die Steuerung der Rotoren für eine Drehung um die eigene Achse wird über die Richtungskontrollpedalen gehandhabt (kollektive Blattverstellung).	Information zu allgemein
Die Umwandlungseffizienz ist durch den Carnotschen Wirkungsgrad (Carnotscher Kreisprozess) prinzipiell nach oben begrenzt (W gewonnene Arbeit, Q eingesetzte Wärmemenge, T1 untere, T2 obere Prozesstemperatur).	Information zu allgemein
Ein angestellter Tragflügel erzeugt bekanntlich auf seiner Oberseite einen Unterdruck und damit verbunden eine erhöhte Geschwindigkeit.	Information zu allgemein
Ein Nachteil der Leeläufer ist die größere Lärmemission und Leistungsschwankungen wegen dem periodischen Gang des Rotorblatts durch die verwirbelte Strömung im Turmnachlauf (oder „Turmschatten“).	Semantische Relation unklar
Etwa in der Mitte der senkrechten Unterwasser-Fachwerkstütze 9 ist der hydraulische Flügelzellenmotor 10 angeflanscht, der den Unterwasserrotor 11 antreibt.	Verallgemeinerbar-keit fraglich
Im Läufer befindet sich eine Wicklung, die direkt oder über einen Widerstand kurzgeschlossen ist.	Information zu allgemein
Jede Etage besteht aus einer Oberboden- und Unterbodenbegrenzung sowie horizontaler Einleitflächenelemente 13, worauf zwölf vertikale Einleitflächenelemente 12 angeordnet sind.	Fachlicher Bezug unklar
Längs des Rotorblatts verwendet man daher üblicherweise verschiedene aerodynamische Profile, dicke Profile im inneren Bereich und dünne Profile im äußeren Bereich.	Information zu allgemein
Mit einer variablen Drehzahl kann sich die Windkraftanlage in einem weiten Drehzahlbereich drehen.	Information zu allgemein
Mithilfe von Frequenzumrichtern wird es somit möglich, auf ein Getriebe zu verzichten und die Generatoren mit der jeweiligen Rotordrehzahl direkt anzutreiben.	Information zu allgemein
Schall mit noch höherer Frequenz wird als Hyperschall bezeichnet, bei Frequenzen unterhalb des für Menschen hörbaren Frequenzbereichs spricht man dagegen von Infraschall.	Information zu allgemein
Stirnradgetriebe werden aus zwei parallel liegenden Zahnrädern konzipiert.	Information zu allgemein
Tripod (Offshore) Die Anlage wird auf einen dreibeinigen Fuß gestellt.	Wörterbuchform
Wegen der gleichmäßigeren Verteilung der Massen- und Luftkräfte über die Rotorkreisfläche sind Dreiblattanlagen dynamisch deutlich ruhiger, was eine geringere Belastung aller Komponenten bewirkt.	Information zu allgemein

Windenergie ist kinetische Energie der Luftteilchen, welche sich mit der Geschwindigkeit bewegen.	Information unvollständig
Windenergieanlagen mit aktiver Stallregelung sind der Versuch, das Konzept der Stallregelung und des netzsynchronen Betriebs ohne teureren Gleich- und Wechselrichter auch auf größere Anlagen bis in den Megawattbereich zu übertragen.	Information zu allgemein
Windgeneratoren, die auf der blossen Mitnahme des Flügels durch die Windströmung basieren, wie etwa der Savonius-Rotor (Ekman-Strömungsmesser), spielen nur eine untergeordnete Rolle.	Information zu allgemein
Windkraftanlagen mit einer horizontale Achse können sich in der Stellung des Rotors zum Turm unterscheiden: beim Leeläufer läuft der Rotor in Windrichtung hinter dem Turm. beim Luvläufer läuft der Rotor in Windrichtung vor dem Turm.	Mehrsatz-Kontext

b) Sub-TKD

Kontext-Kandidat	**Fehlerkategorie**
Abweichungen vom Standard werden als Wärmeanomalien bezeichnet .	Fachlicher Bezug unklar
An Stellen mit hohem induktiven Blindleistungsbedarf dienen zusätzlich übererregte Phasen-schiebermaschinen, die generatorisch induktive Blindleistung liefern, zur Stabilisierung und Entlastung der Versorgungsleitungen, die dann nur Wirkleistung übertragen müssen.	Information zu allgemein
Auch Sendeendröhren größerer Leistung besitzen eine oder mehrere Anschlusskappen .	Information zu allgemein
Außer zum Selbstanlauf verwendet man den Hilfsstrang auch, um das Moment des Motors nach erfolgtem Hochlauf zu erhöhen	Fachlicher Bezug unklar
Bei der indirekten Heizung fließt der Heizstrom durch einen separaten Heizdraht (meistens eine Wolfram - Glühwendel) , der durch Alundum isoliert innerhalb des Kathodenröhrchens liegt .	Fachlicher Bezug unklar
Bei der so genannten Erstarrungstemperatur wird der Abstand so klein , dass sich die Teilchen gegenseitig blockieren und miteinander verstärkt anziehend wechselwirken – sie nehmen eine feste Position in einem dreidimensionalen Gitter ein .	Fachlicher Bezug unklar
Bei der so genannten Hell-Dunkel-Schaltung wird die Differenzspannung zwischen den einzelnen Phasen durch Lampen angezeigt.	Information zu allgemein
Bei fehlender Spannung zwischen Anode und Kathode fließt nur ein minimaler Strom – der sogenannte Anlaufstrom .	Information zu allgemein
Bei Feststoffen unterscheidet man auch nach anderen Merkmalen : kristallin : Ein Feststoff , der seine Form nicht verändert . Seine Bausteine , die Kristalle , weisen eine Fernordnung auf . amorph : Ein Feststoff , der lediglich durch eine Nahordnung ausgezeichnet ist , siehe amorphes Material . Ein amorpher Festkörper ist metastabil .	Mehrsatz-Kontext
Bezogen auf die Erd - Gesamtmasse bestehen etwa 0 , 12 % und bezogen auf die Erdkruste etwa 2 , 9 % aus Wasserstoff .	Semantische Relation unklar

Der Erregerstrom IEO bei der der leerlaufende Generator seine Nennspannung erreicht, heisst Leerlauferregerstrom.	Fachlicher Bezug unklar
Der überkritische Zustand tritt bei Überschreiten des kritischen Punktes auf und ist ein Mischzustand zwischen flüssig und gasförmig .	Fachlicher Bezug unklar
Die Erregermaschine ist eine Außenpolmaschine entsprechend Bild 8.3d.	Fachlicher Bezug unklar
Die gesamte Ständerwicklung besteht aus drei Wicklungssträngen, deren Anfänge und Enden auf das Klemmbrett des Motors geführt sind.	Verallgemeinerbar-keit fraglich
Die im Erdinneren gespeicherte Wärme stammt von der Entstehung des Sonnensystems her .	Terminus unklar
Die Phasendifferenz wird als Polradwinkel J bezeichnet.	Fachlicher Bezug unklar
Die Regelkennlinien beschreiben, wie durch den Erregerstrom die Blindleistung eingestellt werden kann.	Information zu allgemein
Die Richtung des in der Leiterschleife entstehenden Stromes ergibt sich aus dem Lenzschen Gesetz.	Information zu allgemein
Durch Kernfusion werden in der Sonne große Mengen Energie freigesetzt , die als Solarstrahlung (elektromagnetische Strahlung) die Erde erreichen .	Information zu allgemein
Es gibt drei klassische Aggregatzustände : fest : In diesem Zustand behält ein Stoff im Allgemeinen sowohl Form als auch Volumen bei ; siehe Festkörper . flüssig : Hier wird das Volumen beibehalten , aber die Form ist unbeständig und passt sich dem umgebenden Raum an ; siehe Flüssigkeit . gasförmig : Hier entfällt auch die Volumenbeständigkeit , ein Gas füllt den zur Verfügung stehenden Raum vollständig aus ; siehe Gas .	Mehrsatz-Kontext
Große Temperaturunterschiede können unterschiedliche Wärme-dehnung der Maschinenteile bewirken.	Terminus unklar
Größere Schenkelpolläufer bestehen aus geschweissten Radkränzen auf die die Pole aufgeschraubt oder in Schwalbenschwanznuten eingeschoben werden.	Fachlicher Bezug unklar
In den Nuten des Ständers liegt eine Wicklung, die aus mehreren Teilspulen besteht.	Information zu allgemein
In Sternschallung sind die Wicklungsstrangenden W2, U2 und V2 zu einem Sternpunkt verbunden.	Terminus unklar
Indirekt geheizte Bariumoxidkathode – eine Wolfram - Heizwendel wird elektrisch isoliert in ein Nickelröhrchen eingeschoben .	Wörterbuchform
Insgesamt wird beim Wasserstoffbrennen etwa 0 , 73 % der Masse in Energie umgewandelt , was man als Massendefekt bezeichnet .	Fachlicher Bezug unklar
Synchronmotoren werden aber auch zur Lieferung induktiver Bundleistung herangezogen.	Semantische Relation unklar
Temperaturänderungen bewirken ebenso eine Veränderung des Volumens nach den Gesetzen der Wärmeausdehnung .	Terminus unklar
Wolframkathode – der Heizdraht dient gleichzeitig als Elektronenemitter (direkt geheizte Kathode) und besteht wie bei einer Glühlampe aus Wolframdraht .	Wörterbuchform

Anhang 5: Falsche Positive in den russischen Korpora

a) Sub-EKR

Kontext-Kandidat	Fehlerkategorie
«Ездовым циклом» называется состояние двигателя, при котором он проходит режим работы «с обратной связью».	Verallgemeinerbarkeit fraglich
Балка задней подвески состоит из двух продольных рычагов 15 (рис.17) и соединителя 14, которые сварены между собой через усилители.	Verallgemeinerbarkeit fraglich
Барабанные тормозные механизмы различают по типам разжимных устройств.	Information zu allgemein
В радиаторах элементарные каналы, по которым движется охлаждающий воздух, имеют различную форму поперечного сечения: прямоугольную, квадратную, треугольную, полукруглую и т.д.	Terminus unklar
Водяной насос нагнетает жидкость в систему, и основной ее поток проходит по водяной рубашке блока цилиндров от его передней части к задней.	Verallgemeinerbarkeit fraglich
Водяные радиаторы имеют по глубине три-шесть рядов труб.	Information zu allgemein
Воздух из пневмосистемы управления поступает в пневмоцилиндр, перемещает поршень и сжимает пакет,, состоящий из ведущих, ведомых и нажимного тормозных дисков.	Terminus unklar
Впускной распределительный вал вращается в корпусе газораспределительного механизма и приводится зубчатоременной передачей непосредственно от коленчатого вала.	Verallgemeinerbarkeit fraglich
Выборка клапанных зазоров производится автоматически за счет оснащения опорных стоек приводных рычагов гидравлическими компенсаторами.	Information zu allgemein
Гидровакуумный усилитель состоит из вакуумного цилиндра, гидравлического цилиндра и следящего механизма.	Dopplung
Датчик определяет момент прихода поршней цилиндров №1 и №4 в верхнюю мертвую точку на такте сжатия.	Terminus unklar
Датчик представляет собой резистор, который изменяет напряжение сигнала в зависимости от температуры поступающего во впускной коллектор воздуха.	Fehlerhafte Information
Датчик представляет собой термистор, чувствительный к изменению температуры.	Information zu allgemein
Дисковая тормозная система снабжена металлическим диском или ротором (или же сделанным из другого экзотического материала у некоторых гоночных автомобилей).	Fachlicher Bezug unklar
Для лучшего отвода тепла из рабочей зоны часто используют вентилируемые диски.	Information zu allgemein

Для поддержания необходимого давления в системе предусмотрен регулятор давления.	Fachlicher Bezug unklar
Если ранние их варианты полагались на механический привод — сначала тягами, а позднее проложенными между закреплёнными на раме шкивами тросами (вроде тех, которые в наши дни приводят в действие стояночный тормоз) — то начиная со второй половины тридцатых-сороковых годов общепринятыми становятся гидравлические тормозные системы (первые серийные автомобили с ними появились ещё во второй половине двадцатых), в которых тормозные механизмы приводились в действие через длинные системы трубок, заполненных гидравлической жидкостью — изначально изготовлявшейся на основе растительного масла.	Information zu allgemein
К колесным агрегатам относятся (ступицы колес, элементы системы центральной накачки шин (СЦНШ(, ободы колес, – относительная масса ступиц.	Sprachliche Störung
Кислородные датчики, как правило, оборудуются нагревателем, который нагревает чувствительный циркониевый элемент.	Dopplung
Коленчатый вал вращается в пяти коренных подшипниковых опорах.	Semantische Relation unklar
Коленчатый вал вращается в пяти оборудованных подшипниками скольжения коренных опорах.	Semantische Relation unklar
Конструкция шарнира обеспечивает одинаковые угловые скорости ведущего и ведомого валов независимо от угла между ними.	Terminus unklar
Крыльчатка водяного центробежного насоса и вентилятор приводятся в действие приводным усторойством.	Information zu allgemein
На часть автомобилей устанавливают дисковые тормозные механизмы задних колес, аналогичные по конструкции механизмам передних колес, но меньшего размера.	Information zu allgemein
Наибольший эффект регулирования тормозных сил автомобиля обеспечивает регулятор тормозных сил в пневмоприводе тормозов, имеющих упругую связь с задним мостом, который учитывает статическое и динамическое перераспределение веса автомобиля.	Information zu allgemein
Оба распределительных вала вращаются в головке блока и приводятся от коленчатого вала резиновым зубчатым ремнем, одновременно служащим также для привода водяного насоса.	Verallgemeinerbar-keit fraglich
Отлитые из алюминия поршни снабжены тремя поршневыми кольцами: двумя компрессионными и одним (нижним) маслосъемным.	Fachlicher Bezug unklar
Положением дроссельной заслонки определяется, сколько воздуха попадёт во впускной коллектор при данном перепаде давлений между атмосферным давлением и давлением во впускном коллекторе.	Dopplung

Поэтому закрытые системы всегда оборудуют так называемыми паровыми клапанами, ограничивающими давление для предотвращения возможных разрушений рубашек цилиндров и блоков, трубопроводов, охладителей, расширительных баков, а также нарушения плотности соединений, в частности, уплотнений гильз и втулок цилиндров.	Fachlicher Bezug unklar
При вращении крыльчатки насоса охлаждающая жидкость нагнетается в водораспределительную трубку 12, расположенную в головке блока.	Information zu allgemein
При вращении крыльчатки насоса охлаждающая жидкость нагнетается в водораспределительную трубку, расположенную в головке блока.	Information zu allgemein
Привод водяного насоса и вентилятора осуществлен от шкива коленчатого вала при помощи клиноременной передачи, состоящей из двух ремней.	Information zu allgemein
Привод впускных и выпускных клапанов осуществляется непосредственно от эксцентриковых кулачков соответствующих распределительных валов через оборудованные гидравлическими компенсаторами клапанных зазоров толкатели.	Fachlicher Bezug unklar
Привод же выпускного распределительного вала осуществляется посредством шестеренчатой передачи от впускного вала.	Information zu allgemein
Привод осуществляется на передние колеса посредством закрепленной слева на двигателе трансмиссионной сборки.	Information zu allgemein
Принцип измерения расхода воздуха датчиком этого типа основан на изменении теплопереноса от поверхности пленки к потоку протекающего через датчик воздуха.	Terminus unklar
Работа усилителя основана на использовании разрежения во впускной трубе работающего двигателя.	Fachlicher Bezug unklar
Распределительный вал вращается в постели, укрепленной сверху головки цилиндров, и приводится во вращение от коленчатого вала посредством резинового зубчатого ремня (тот же ремень используется для привода водяного насоса).	Fachlicher Bezug unklar
Распределительный вал приводится от коленчатого вала посредством зубчатого ремня и вращается непосредственно в головке блока.	Fachlicher Bezug unklar
РЕГУЛЯТОР ДАВЛЕНИЯ Это устройство регулирует давление в гидравлическом приводе тормозных механизмов задних колес в зависимости от нагрузки на заднюю ось автомобиля.	Wörterbuchform
PCM управляет подачей топлива в цилиндры двигателя, расходом воздуха на всех режиме холостого хода двигателя и углом опережения зажигания.	Dopplung
PCM, в свою очередь, подстраивает сервопривод регулятора оборотов холостого хода таким образом, чтобы компенсировать снижение оборотов холостого хода, вызываемые нагрузкой насоса гидроусилителя рулевого управления.	Fachlicher Bezug unklar
Рулевое управление — реечное, с гидроусилителем, рулевой механизм расположен за двигателем.	Verallgemeinerbar-keit fraglich

Силовой транзистор системы зажигания, установленный в первичной цепи системы, включается и выключается для изменения тока в первичной цепи катушки зажигания.	Dopplung
Система воздушного охлаждения автомобильных и тракторных двигателей состоит из ряда элементов, регулирующих ее работу и поддерживающих заданное тепловое состояние двигателя.	Information zu allgemein
Смазывание кулачков распределительного вала и клапанных компонентов, равно как и прочих внутренних компонентов двигателя осуществляется методом разбрызгивания.	Information zu allgemein
Стойки и пластина образуют каркас радиатора.	Fachlicher Bezug unklar
Стояночная тормозная система приводится в действие с помощью рычага стояночного тормоза.	Dopplung
Теплота, отводимая в охлаждающую жидкость внутреннего контура (воду, ТОСОЛ,антифриз) и смазочное масло, а также теплота, воспринимаемая при охлаждении наддувочного воздуха, передается охлаждающему теплоносителю в соответствующих охладителях.	Terminus unklar
Торможение осуществляется за счет раздвигания колодок и прижатия их к барабану с помощью рабочего цилиндра, в который поступает тормозная жидкость при включении педали тормоза.	Terminus unklar
Торможение передних колес осуществляется за счет сил трения, возникающих между закрепленным на спице колеса чугунным тормозным диском и прижимаемыми к нему с двух сторон тормозными колодками с накладками.	Information zu allgemein
Тормозная система оснащена штатным вакуумным усилителем.	Information zu allgemein
Тормозная система состоит из главного тормозного цилиндра, вакуумного усилителя тормозов и рабочих тормозных механизмов на передних и задних колесах.	Dopplung
Тормозной механизм передних колес автомобиля ЗАЗ-966 (рис.) состоит из тормозного щитка, прикрепленного болтами к поворотной цапфе, двух рабочих колесных цилиндров (верхний диаметром 22 мм, нижний - 19 мм), каждый из которых приводит в действие одну из двух тормозных колодок.	Dopplung
Тормозные системы должны обеспечивать равномерное распределение тормозных сил между колесами одного моста.	Information zu allgemein
Увеличенная толщина вентилируемого диска позволяет разместить между поверхностями трения ребра жесткости, которые обеспечивают принудительную циркуляцию воздуха.	Fachlicher Bezug unklar
Установка клапанных зазоров осуществляется вручную путем подбора шайб требуемой толщины.	Terminus unklar
Фрикционная лента представляет собой стальную ленту 10 с приклепанными асбестовыми накладками толщиной 8 мм и приваренными кронштейнами.	Dopplung
Шарнир состоит из двух вилок, в криволинейных канавках которых расположены четыре шарика.	Verallgemeinerbar-keit fraglich

Эксцентриковые кулачки распределительного вала воздействуют на рычаги привода клапанов, рабочие концы которых упираются в торцы продетых в направляющие втулки и оснащенных маслоотражательными колпачками стержней подпружиненных клапанов.	Prozess unklar
Электрогидравлический рулевой насос установлен непосредственно на рулевом механизме и не требует обслуживания.	Verallgemeinerbar-keit fraglich
Электронный блок управления двигателем подает контрольное напряжение на датчик положения дроссельной заслонки (TPS) и затем измеряет напряжение в цепи сигнала датчика.	Fachlicher Bezug unklar
Электронный блок управления, в соответствии с величиной электродвижущей силы кислородного датчика определяет степень отклонения состава смеси от стехиометрического и, в соответствии с этим, подстраивает необходимое количество впрыскиваемого топлива путем изменения продолжительности сигнала управления форсунками.	Fachlicher Bezug unklar
Эффект самоподвода колодок обеспечивается манжетой поршня (есть и более сложные системы подвода колодок в дисковых тормозах).	Fachlicher Bezug unklar
Эффективность тормозного пути определяется по определенной оценке тормозного пути или временем движения автомобиля до полной остановки.	Information zu allgemein

b) Sub-TKR

Kontext-Kandidat	**Fehlerkategorie**
Биогаз - биотопливная замена природного газа.	Wörterbuchform
Большинство установок солнечного теплоснабжения оборудовано солнечным коллектором.	Information zu allgemein
Бункер для панировочных сухарей установлен над столом перед бункером для фарша и имеет коническую съёмную воронку.	Verallgemeinerbar-keit fraglich
В кабине устанавливают рубильник или магнитный пускатель для включения источника сварочного тока.	Verallgemeinerbar-keit fraglich
В полуавтоматах сварочных не применяются электроды, их заменила проволока, по которой подается газ для сварки, что и является главным их отличием от всех других типов аппаратов.	Fachlicher Bezug unklar
В системах открытых камер газообразные продукты электролиза покидают рабочий объём, что затрудняет калькуляцию баланса полученной/затраченной энергии.	Fachlicher Bezug unklar
В состав наземного специального комплекса навигационно-информационной системы входят интегрированная телематическая система транспортных коридоров и система обеспечения потребителей высокоточной навигационной информацией.	Verallgemeinerbar-keit unklar

В экспериментах с закрытыми камерами продукты электролиза утилизируются, например, путем каталитической рекомбинации в специальных частях системы.	Fachlicher Bezug unklar
Введение добавочного металла осуществляется посредством расплавления сварочным источником тепла специальных сварочных материалов.	Terminus unklar
Входящая в состав МАКСМ Международная навигационно-информационная система предназначена для обеспечения потребителей различных государств навигационной информацией, получаемой космическими навигационными системами, в интересах проведения эвакуационных мероприятий, а также решения других социально-экономических задач, в том числе и информационное и телекоммуникационное обеспечение.	Verallgemeinerbar-keit fraglich
Выработка электроэнергии осуществляется двумя синхронными генераторами типа TAG 8123, с воздушной косвенной системой охлаждения, номинальной мощностью 35 МВт каждый, напряжением 10,5 кВ, частотой 50 Гц, сопряженными с паровыми турбинами.	Verallgemeinerbar-keit fraglich
Газоснабжение котельной осуществляется от газопровода среднего давления.	Information zu allgemein
Гребёнка совершает сложное движение, состоящее из поступательного движения перпендикулярно оси колеса и возвратно-поступательного движения (на анимации не показано), параллельного оси колеса для снятия стружки по всей ширине его обода.	Prozess unklar
Деаэрация воды в установках ВДПУ происходит в результате кипения воды на керамических насадках за счет разряжения, создаваемого в деаэрационной колонке.	Prozess unklar
Для образования угля необходимо обильное накопление растительной массы.	Information zu allgemein
Для того, чтобы зафиксировать результаты, полученные в ходе проверки, ведется специальный журнал учета.	Fachlicher Bezug unklar
Для транспортных устройств различают массовую Q (кг/ч) и объемную V (м3/ч) производительность.	Semantische Relation unklar
Дополнительные параметры - состав и толщина покрытия электрода, положение электрода и положение изделия.	Fachlicher Bezug unklar
Другие изотопы генерируются в охлаждающей воде, включая изотопы аргона, фтора, водорода и кислорода; ряд других изотопов образуется в материалах конструкций, окружающих ядро реактора.	Information zu allgemein
Испытание керосином - применяется для определения плотности сварных швов.	Wörterbuchform
Керамические "светлые" инфракрасные излучатели работают на основе беспламенного сгорания газо-воздушной смеси на поверхности керамической плитки, что приводит к ее нагреванию до температуры 850…900ºС.	Fachlicher Bezug unklar
Кислородный баллон представляет собой стальной цилиндр со сферическим днищем и горловиной для крепления запорного вентиля.	Fachlicher Bezug unklar

Комплект источников питания состоит из понижающего трансформатора и дросселя (регулятора активной катушки).	Verallgemeinerbar-keit fraglich
Конструкция дросселя со ступенчатым регулированием сварочного тока позволяет изменять величину сварочного тока при помощи передвигающегося контакта путем включения определенного количества витков обмотки.	Prozess unklar
Космическая астрофизика - наблюдения отдаленных галактических и внегалактических объектов с космических аппаратов с помощью космических телескопов во всех диапазонах электромагнитного излучения и с помощью приемников космических лучей (см. здесь).	Wörterbuchform
Метрическая резьба — с шагом и основными параметрами резьбы в долях метра.	Wörterbuchform
Модульная технология создания космической платформы позволяет с небольшими затратами и в короткие сроки адаптировать возможности платформы к применению в составе космических аппаратов разного типа с разной целевой аппаратурой.	Terminus unklar
Нагрев воды на нужды горячего водоснабжения осуществляется в водоводяном подогревателе 1x57x2000xP-4, циркуляция воды – насосами горячего водоснабжения (один рабочий и один резервный).	Information zu allgemein
Нагрев воды осуществляется газовыми котлами серии "Десна", установленными на раме, прикрепленной к полу контейнера.	Verallgemeinerbar-keit fraglich
НТД включает стандарты, технические условия, чертежи изделий, технологические карты и производственные инструкции.	Fachlicher Bezug unklar
Основные параметры- это величина и полярность тока, диаметр электрода, напряжение на дуге, скорость сварки.	Fachlicher Bezug unklar
Основными требованиями для всех типов электродов являются: обеспечение стабильного горения дуги и хорошего формирования шва; получение металла сварного шва заданного химического состава; спокойное и равномерное расплавление электродного стержня и покрытия; минимальное разбрызгивание электродного металла и высокая производительность сварки; легкая отделимость шлака и достаточная прочность покрытий; сохранение физико-химических и технологических свойств электродов в течение определенного промежутка времени; минимальная токсичность при изготовлении и при сварке.	Semantische Relation unklar
Отпуск тепла от турбины осуществляется через промежуточный контур водой под давлением, что является дополнительным барьером от проникновения радиоактивности к потребителям тепла.	Prozess unklar
Перспективы использования возобновляемых источников энергии связаны с их экологической чистотой, низкой стоимостью эксплуатации и ожидаемым топливным дефицитом в традиционной энергетике.	Information zu allgemein
планетология – исследование планет и малых тел Солнечной системы;	Wörterbuchform

Поверхностная резка - режущая струя направлена под очень малым углом к поверхности металла (почти параллельно ей) и обеспечивают грубую ее строжку или обдирку.	Wörterbuchform
Под головной материаловедческой организацией понимается головная отраслевая материаловедческая организация, если нет других специальных указаний в тексте.	Verallgemeinerbarkeit fraglich
Подход Лауэ отличается от подхода Брегга тем, что в нем не выделяется никакой конкретный способ разбиения кристалла на атомные плоскости и не используется специальное предположение относительно зеркального характера отражения.	Terminus unklar
Посты подразделяются на стационарные и передвижные.	Fachlicher Bezug unklar
Преобразователем напряжения называют главный составляющий элемент, который присутствует в устройстве сварочных аппаратов.	Information zu allgemein
Привод взбивалки состоит из электродвигателя 8, клиноременного вариатора, цилиндрического 4 и конического 3 редукторов и планетарного механизма 11 с крышкой.	Verallgemeinerbar-keit fraglich
Привод состоит из электродвигателя с клиноременной передачей.	Fachlicher Bezug unklar
Приводной механизм котлетоформировочной машины состоит из электродвигателя, червячного редуктора и зубчатой цилиндрической передачи.	Verallgemeinerbar-keit fraglich
Проволока служит для подвода тока и пополнения сварочной ванны расплавленным металлом.	Prozess unklar
Промежутки времени между противостояниями Марса и Земли составляют около 25 месяцев, но расстояния между ними в эти моменты оказываются раз от раза различными.	Semantische Relation unklar
Прямозубые колёса — самый распространённый вид зубчатых колёс.	Wörterbuchform
Рабочей камерой служит вращающийся формирующий стол, который имеет ячейки круглой или овальной формы, в который установлены поршни.	Verallgemeinerbar-keit fraglich
Резка кислородным копьем - копье образуется тонкостенной стальной трубкой, присоединенной к рукоятке и свободным концом прижатой к прожигаемому металлу.	Wörterbuchform
С помощью манипуляторов можно регулировать скорость, с которой вращается планшайба, а также скорость сварного узла, который крепится на планшайбе.	Prozess unklar
С помощью управляющих воздействий обеспечивают требуемые показатели качества и их стабильность.	Information zu allgemein
Секторное колесо представляет собой часть обычного колеса любого типа.	Information zu allgemein

Система обеспечения потребителей высокоточной навигационной информацией предназначена для формирования навигационно-информационного пространства, в пределах которого для неограниченного числа мобильных и стационарных объектов, находящихся в любой точке земной, водной поверхности и воздушной среды, обеспечивается возможность автоматического высокоточного определения координат местоположения по сигналам спутниковых навигационных систем ГЛОНАСС/GPS/Galileo.	Verallgemeinerbar-keit fraglich
Стационарные рубительные машины используют в качестве привода электродвигатели различной мощности и применяются чаще всего в линиях по производству строительных материалов из древесины в целях утилизации её отходов.	Prozess unklar
Стопорение — предотвращение самоотвинчивания.	Wörterbuchform
Топливом для котельных служит природный газ, резервное топливо не предусматривается.	Information zu allgemein
Турбина теплофикационного типа, предназначена для выработки тепла и привода генератора как источника электроэнергии.	Information zu allgemein
Удаление воздуха осуществляется через дефлекторы.	Fachlicher Bezug unklar
Фильтр состоит из корпуса 7 с фильтрующей набивкой 6, под-дона 1 с маслоразделителем, диафрагмы 2, стакана диафрагмы 5 и пружины 4.	Terminus unklar
Циркуляция воды в системе отопления и вентиляции осуществляется сетевыми насосами (два рабочих и один резервный).	Information zu allgemein

Anhang 6: Unbekannte KRC-Kandidaten in den russischen Subkorpora

a) Sub-EKR

Для высокофорсированных дизелей в трубчатоленточных масляных радиаторах применяют внутренние вставки - завихрители, что позволяет при тех же размерах сердцевины в 2,7-3 раза повысить теплоотдачу по сравнению с гладкотрубчатыми масляными радиаторами.
Тормозные шланги осуществляют подвижные связи между подвижными и неподвижными частями автомобиля.
Разобщительный кран служит для отключения магистрали от при-цепа и устанавливается перед соединительной головкой.
Стояночный (ручной) тормоз имеет механический тросовый привод на колодки задних тормозов.
Дисковый тормозной механизм составляют: тормозной диск, поршень с манжетом, поршни с манжетом, тормозная колодка.
Один контур гидропривода обеспечивает работу правого переднего и левого заднего тормозных механизмов, другого – левого переднего и правого заднего.
В соответствии с сигналом датчика температуры воздуха во впускном коллекторе электронный блок управления двигателем будет корректировать необходимую подачу топлива (базовое время открытого состояния топливной форсунки).

b) Sub-TKR

Промышленная продукция - конечный результат деятельности промышленных предприятий.
Необходимость отпуска вызывается тем, что в закаленной детали появляется не только твердость но и большие внутренние напряжения.
Сварка лежачим электродом производится покрытым электродом различной длины, но не более 1200 миллиметров, и диаметром до 8 миллиметров.
Геотермальная энергия может быть использована двумя основными способами — для выработки электроэнергии и для обогрева домов, учреждений и промышленных предприятии[1]
Трещины (продольные и поперечные, по шву и околошовные) создают несплошность материала для силового потока и вследствие этого местную концентрацию напряжений с резким падением динамической и вибрационной прочности конструкции.
Передаточное число используется при расчётах геометрических параметров зубчатых передач.
Оценку качества промышленной продукции производят путём её контроля, т.е. проверки соответствия контролируемых показателей заданным требованиям или нормам, установленным нормативно-технической документацией (НТД).

[1] sic!

Anhang 7: Unbekannte KRC-Kandidaten in den deutschen Subkorpora

a) Sub-EKD

Nach dem zweiten Axiom von Newton ist die Auftriebskraft proportional zur Masse der bewegten Luft und deren Geschwindigkeit bzw. Beschleunigung (F=m•a).
Durch die Blattverstellung kann der Anstellwinkel des Rotorblatts in bezug zur Anströmung kontinuierlich so verändert werden, daß die Leistungsaufnahme des Rotors gesteuert werden kann.[1]
Die ältesten, bekannten Windkraftanlagen der Welt (seit 1700 v. Chr. Gebaut) sind vertikale Windkraftanlagen und sind im Iran und Afghanistan zu finden.[2]
Ultraschall kann von verschiedenen Tieren (Fledermäuse, Wale, Delfine) oder durch technische Anordnungen erzeugt werden.
Anlagen mit aktiver Stallregelung und Pitchregelung können die Rotorblätter aus dem Wind drehen und aerodynamisch abbremsen.
Wenn die Läuferdrehzahl (Drehzahl des vom Getriebe übersetzten Rotors) über der Drehfelddrehzahl liegt, dann ist die Asynchronmaschine im Generatormodus, und wenn sie darunter liegt, dann arbeitet sie als Motor.

b) Sub-TKD

Die Widerstandsdreiecke veranschaulichen die Größenverhältnisse von Wirkwiderstand RL und induktivem Blindwiderstand XL beim Anlauf und im Nennbetrieb.
Der Stirling- Motor findet vor allem in der Landwirtschaft einige Verbreitung, weil man ihn mit jedem Brennstoff, also auch mit Holz oder Stroh, betreiben kann.[3]
Ammoniak ist der Grundstoff für die Herstellung aller anderen industriell hergestellten stickstoffhaltigen Verbindungen.
In der Brennstoffzelle wird aber die Energie in einem chemischen Treibstoff nicht über den Umweg der Verbrennung erzeugt, sondern katalytisch direkt in elektrischen Strom verwandelt, der dann die Elektromotoren antreibt.
Zur Beheizung von Gebäuden werden im unteren Leistungsbereich meist Elektro-Kompressions- Wärmepumpen verwendet, bei höheren Leistungen auch Gasmotorwärmepumpen.[4]
Infolge des Stromverdrängungseffektes kommt es bei den Stromverdrängungsläufern zu einer Erhöhung des Widerstandes der Läuferwicklung beim Anlauf.
Handels- U- Boote kamen lediglich in den beiden Weltkriegen zum Einsatz, um feindliche Seeblockaden zu umgehen, mit neutralen Staaten Handel zu treiben und dabei kriegswichtige Güter zu beschaffen.[5]
Biomasse eignet sich auch zur Herstellung von Wasserstoff in einer Wasserstoffwirtschaft.

[1] sic!
[2] sic!
[3] sic!
[4] sic!
[5] sic!

Die EEG- Quote entspricht dem Anteil EEG- geförderter Strommenge an dem nicht-privilegierten Letztverbrauch.[6]
Der Dreiphasen-Induktionsmotor mit Kurzschlussläufer ist der am weitesten verbreitete Drehstrommotor.

[6] sic!

Anhang 8: URLs der Texte in den russischen Subkorpora

a) Sub-EKR

http://roman.by/r-32637.html
http://silich.ru/so-dvs
http://opel-astra.5go.ru/
http://www.hyundai-sto.ru/santa-fe/194.htm
http://mazda626.ru/?Instrukciya:Tormoznaya_sistema:Diskovyi_tormoznoi_mehanizm_zad nei_osi
http://roman.by/r-58236.html
http://niva-chevy.narod.ru/4x4/8-43.html
http://driverguide.com.ua/school/1-klass/brake-system/Vse-stranitsy.html
http://www.pro-men.ru/main/auto/FAQ/gsws964w
http://mirprom.ru/Spravka/Spravka.aspx?id=1176
http://www.hyundai-sto.ru/santa-fe/197.htm
http://www.navigator.mn/hyundai-santa-fe/dvigatel/toplivnaya-sistema/sistema-upravleniya-raspredelennim-vpriskom-topliva.html
http://focusello.narod.ru/fok2/9-13.html
http://stroy-technics.ru/article/tormoznye-sistemy-avtogreiderov
http://lanosoid.narod.ru/html/9-1c_l.html
http://automn.ru/vaz-2107/autocategory-4319-10.m_id-4318.html
http://ru.wikipedia.org/wiki/%D0%A2%D0%BE%D1%80%D0%BC%D0%BE%D0%B7%D0%BD%D0%B0%D1%8F_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0
http://www.navigator.mn/hyundai-sonata/tormoznaya-sistema.html
http://www.navigator.mn/hyundai-santa-fe/dvigatel/toplivnaya-sistema/poisk-neispravnostey-po-diagnosticheskim-kodam.html
http://pressa.irk.ru/number1/2010/34/023001.html
http://chiptuning.zp.ua/index.php/xolostoi-xod
http://www.chipmaster.ru/articles/tormozsys.htm
http://www.brestauto.com/tormozarticle.htm
http://avtosostav.ru/publ/1-1-0-4
http://www.hyundai-sto.ru/sonata/177.htm
http://www.autocentre.ua/ac/Service/Exploitation/12391.html
http://www.varshavka45.ru/lib/auto1000/040.htm
http://autoend.ru/Golf1/Running/Front_suspensionS_txt.htm
http://www.faqnissan.ru/faq/004/10431.html
http://www.tuning100.ru/catalog/cat2.html
http://www.autoprospect.ru/mercedes-benz/e-class/6-4-3-sistema-zazhiganiya.html
http://www.varshavka45.ru/lib/auto1000/038.htm
http://focusman.ru/p152.html
http://autodela.ru/article/view/3435
http://roman.by/r-67015.html
http://tormozsystem.pp.ru/page1.html

b) Sub-TKR

http://revolution.allbest.ru/manufacture/00325523_0.html
http://revolution.allbest.ru/manufacture/00001146_0.html
http://revolution.allbest.ru/manufacture/00167955_0.html
http://www.mistichno.ru/96-nejtronnoe-oruzhie.html
http://revolution.allbest.ru/manufacture/00299446_0.html
http://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B7%D1%8C%D0%B1%D0%B0
http://ru.wikipedia.org/wiki/%D0%A8%D0%B5%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8F
http://any-book.org/download/10224.html
http://dic.academic.ru/dic.nsf/ruwiki/952598
http://www.tayur-kotly.ru/prod_discript.php
http://www.razym.com/news/new_58.html
http://www.energetica.ru/rus/m_e/peb_rus/m_e_peb_ru.htm
http://www.tayur-kotly.ru/boiler_house_11.php
http://dic.academic.ru/dic.nsf/ntes/1832/%D0%98%D0%A1%D0%9A%D0%9E%D0%9F%D0%90%D0%95%D0%9C%D0%9E%D0%95
http://www.heuristic.su/effects/catalog/est/byId/description/64/index.html
http://autorelease.ru/articles/automobile/345-biotoplivo-problemy-i-perspektivy.html
http://www.federalspace.ru/main.php?id=11&did=436
http://dic.academic.ru/dic.nsf/ruwiki/1015600
http://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%82%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5
http://www.federalspace.ru/main.php?id=11&did=500
http://www.federalspace.ru/main.php?id=2&nid=8034
http://ru.wikipedia.org/wiki/%D0%A5%D0%BE%D0%BB%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D1%82%D0%B5%D1%80%D0%BC%D0%BE%D1%8F%D0%B4%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D1%81%D0%B8%D0%BD%D1%82%D0%B5%D0%B7
http://ru.wikipedia.org/wiki/%D0%90%D0%BB%D1%8C%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%B0%D1%8F_%D1%8D%D0%BD%D0%B5%D1%80%D0%B3%D0%B5%D1%82%D0%B8%D0%BA%D0%B0
http://dic.academic.ru/dic.nsf/ruwiki/50089
http://dic.academic.ru/dic.nsf/ruwiki/71752
http://dic.academic.ru/dic.nsf/ruwiki/885921
http://www.stroy-armatura.ru/index.php?mod=text&uitxt=846
http://www.redmotor.ru/965all/17.html
http://ru.wikipedia.org/wiki/%D0%97%D1%83%D0%B1%D1%87%D0%B0%D1%82%D0%B0%D1%8F_%D0%BF%D0%B5%D1%80%D0%B5%D0%B4%D0%B0%D1%87%D0%B0
http://www.ecoteco.ru/id569/
http://www.federalspace.ru/main.php?id=11&did=435
http://www.federalspace.ru/main.php?id=25
http://www.tayur-kotly.ru/ziosab.php
http://revolution.allbest.ru/manufacture/00340960.html
http://www.skonline.ru/digest/40935.html?page=3
http://www.skonline.ru/digest/40935.html

Anhang 9: URLs der Texte in den deutschen Subkorpora

a) Sub-EKD

http://www.kaufen.com/Preisvergleich/windturbine
http://cgi.ebay.de/Darrieus-Rotor-Windkraftanlage-bauen-Technik-3000-S-/110544662591?pt=DE_Heimwerker_Bio_Wasser_Windenergie&hash=item19bcf9b03f
http://cgi.ebay.de/Rotor-Rotorblatter-2-1md-Nabe-Windkraft-Wind-/180492666076?pt=DE_Heimwerker_Bio_Wasser_Windenergie&hash=item2a063384dc
http://de.wdcs.org/schulprojekt/al_lautemeere.htm
http://de.wikipedia.org/wiki/Windenergie
http://www.freepatentsonline.com/EP1177381.html
http://elite.tugraz.at/Jungbauer/3.htm
http://energy-age.de/unternehmen.php
http://fd0.org/projekte/energiewende/windkraftwerk.php
http://members.tripod.com/wasserstoff_ag/wind/der_aufbau_einer_windanlage.html
http://offshore-wind.net/
http://psychophysischer-terror.de.tl/Strahlenwaffen.htm
http://ryabenkiy.com/wind/technik/
http://www.abc-alternative-energy.de/windenergie/wie-funktioniert-windrad.html
http://www.abipur.de/hausaufgaben/neu/detail/stat/253643743.html
http://www.alternative-energie-24.de/windenergie.html
http://www.dualsphere.com/energiedach/windkraftanlagen.php
http://www.energyprofi.com/jo/Alles-ueber-Wind-wind-energy-und-Windenergien.html
http://www.exergia.de/htm.Wind.Power/Wind.Power.all_de.htm
http://www.experten-tricks.at/997,die-funktionsweise-eines-asynchrongenerators/
http://www.hamburgerfirmen.de/firmennews.php?quelle=&suche=&nstart=441
http://www.hans-peterweber.de/windkraftanlagen-entscheidungen.html
http://www.hdt-essen.de/htd/verein/themenbereiche/windenergie.html
http://www.heliport.de/lexika/hubschrauber-physiklexikon/
http://www.iem.ing.tu-bs.de/paper/2002/turki_02.htm
http://www.ipu-bs.de/index.php?id=98&type=98
http://www.irb.fraunhofer.de/denkmalpflege/monulit.jsp?s=Rotor
http://www.kalkriese.de/Windenergieanlage.html
http://www.lerntippsammlung.de/Generator.html
http://www.nicht-fossil.de/64/repeller.htm
http://www.oekoadressen.de/kat/28/windkraft/
http://www.oeko-energie.de/produkte/windkraft/index.php
http://www.oel-alle.de/windkraftanlagen-privathaushalte-variantenueberblick/319/
http://www.pcs-converter.de/index.php?article_id=30&clang=0
http://www.renewable-energy24.eu/de/windkraft.html
http://www.robertmelchner.de/Windenergie.htm

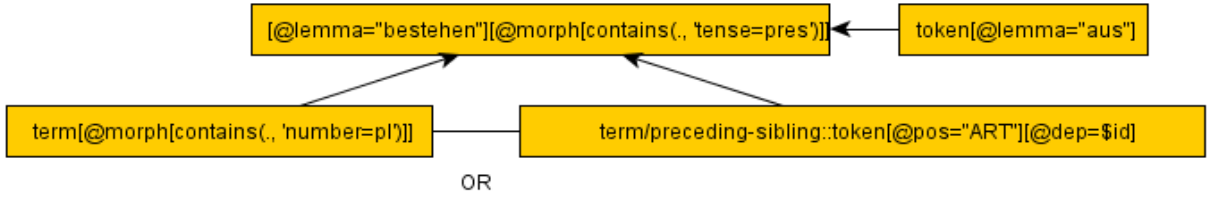
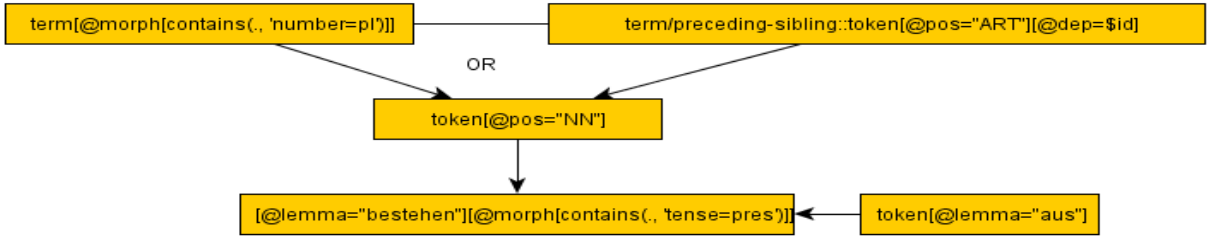
http://www.ruhdorfer.de/diplom/windkraft.html
http://www.segelladen.de/Inhalt-untergruppen3/windgeneratoren.htm
http://www.strom-aus-windkraft.de/Kleinanzeigen/index.php?id=25
http://www.talentfactory.dk/de/tour/design/horver.htm
http://www.techniklexikon.net/d/energietechnik/energietechnik.htm
http://www.techniklexikon.net/d/windenergie/windenergie.htm
http://www.tf-power.de/0430039b0a11a8725/0430039b0a11b2d33/index.php
http://www.umweltdatenbank.de/lexikon/windenergie.htm
http://www.uni-kiel.de/Geographie/lehrv_online/Pellworm/pellwormcd/boris/boris.html
http://www.uni-marburg.de/fb03/politikwissenschaft/energieundsicherheit/traeger/35bwindenergie
http://www.uni-muenster.de/Physik.TD/leistungsbeiwert_windkraftanlage.html
http://www.vde.com/de/fg/ETG/Arbeitsgebiete/V1/Aktuelles/Oeffentlich/Seiten/Windenergie.aspx
http://www.wind-energie.de/de/technik/konstruktiver-aufbau/kupplung/?type=97
http://www.windenergie-nord.de/
http://www.windkraft-gloett.de/
http://www.windkraft-journal.de/kooperationen.html
http://www.wissenschaft-frankreich.de/publikationen/wissenschaft_frankreich/nummer/files/174.htm
http://www.wotys.de/windkraft.htm
http://www-eev.uni-paderborn.de/forschung/windenergie/windparkplanung
http://velagear.com/grm/history.html
http://www.tcw.de/news/view/549

b) Sub-TKD

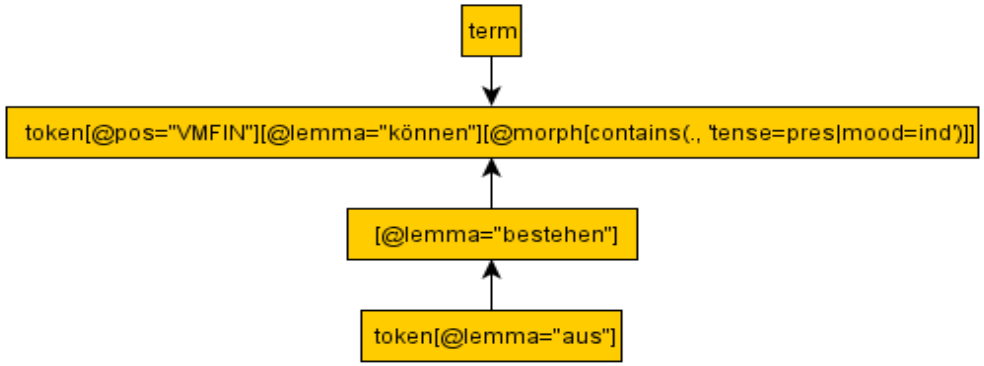
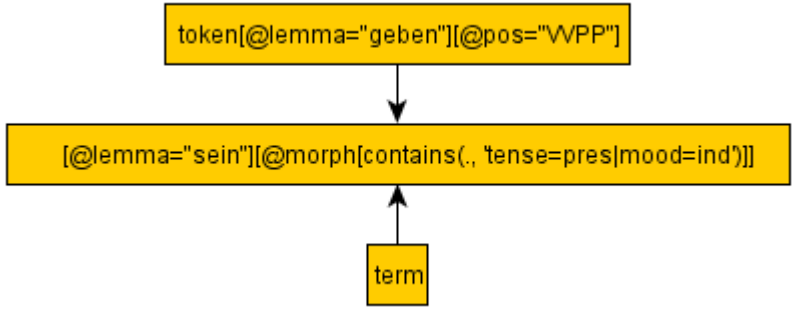
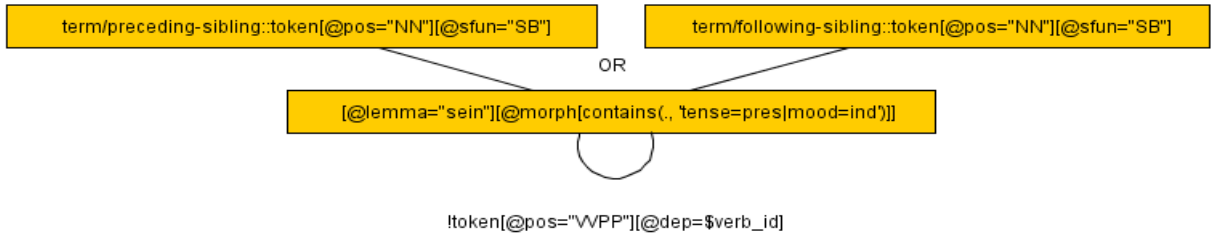
http://www.elektrosoftware.com/extras/ETHMotoren.pdf
http://de.wikipedia.org/wiki/Elektronenr%C3%B6hre
http://de.wikipedia.org/wiki/U-Boot
http://de.wikipedia.org/wiki/Aggregatzustand
http://de.wikipedia.org/wiki/Normung
http://de.wikipedia.org/wiki/Energiequelle
http://de.wikipedia.org/wiki/Wasserstoff
http://de.wikipedia.org/wiki/Tiefe_Geothermie
http://de.wikipedia.org/wiki/Gas
http://de.wikipedia.org/wiki/Ammoniak
http://de.wikipedia.org/wiki/Stromerzeugung
http://de.wikipedia.org/wiki/Erfindungsh%C3%B6he
http://de.wikipedia.org/wiki/Regenerativ
http://de.wikipedia.org/wiki/Erneuerbare-Energien-Gesetz
http://www.buch-der-synergie.de/c_neu_html/c_04_37_sonne_hochtemperatur_stirling.htm
http://de.wikipedia.org/wiki/Niederlande
http://www.maschinenmarkt.vogel.de/themenkanaele/konstruktion/antriebstechnik_steuerungstechnik/articles/195112/
http://www.treefinder.de/Ideas/Waermekraftmaschine.pdf
http://de.wikipedia.org/wiki/Kostendeckende_Einspeiseverg%C3%BCtung_(Schweiz)

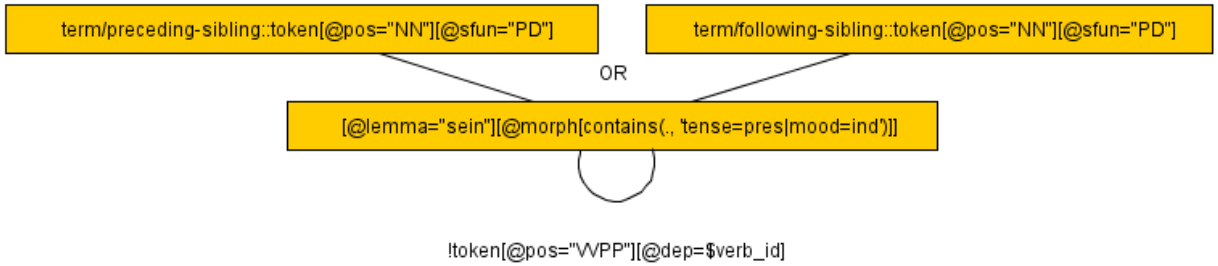
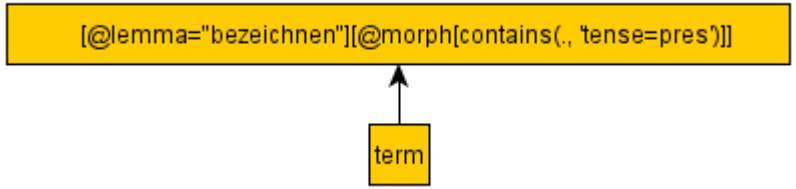
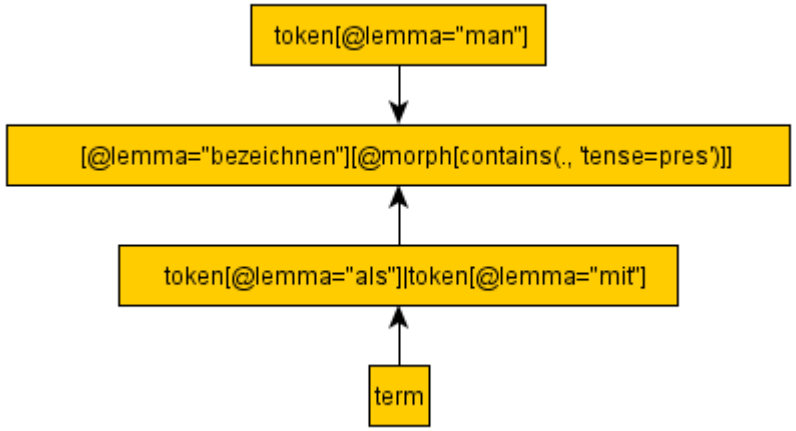
http://de.wikipedia.org/wiki/Pariser_Verbands%C3%BCbereinkunft_zum_Schutz_des_gewerblichen_Eigentums
http://de.wikipedia.org/wiki/Schmier%C3%B6l
http://de.wikipedia.org/wiki/W%C3%A4rmepumpenheizung
http://de.wikipedia.org/wiki/Priorit%C3%A4t
http://www.imab.tu-bs.de/download/WS0809_EU_Kap7u8_v5.pdf

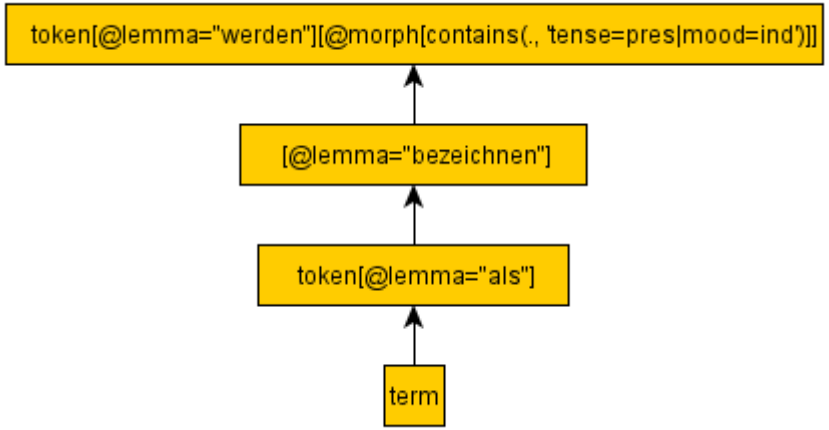
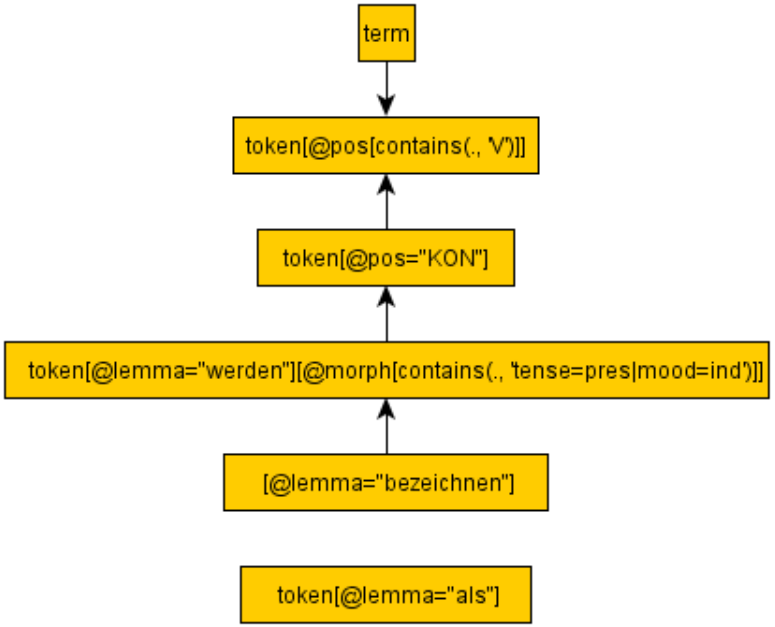
Anhang 10: Genutzte Extraktionsmuster und durch sie extrahierte KRCs – Deutsch

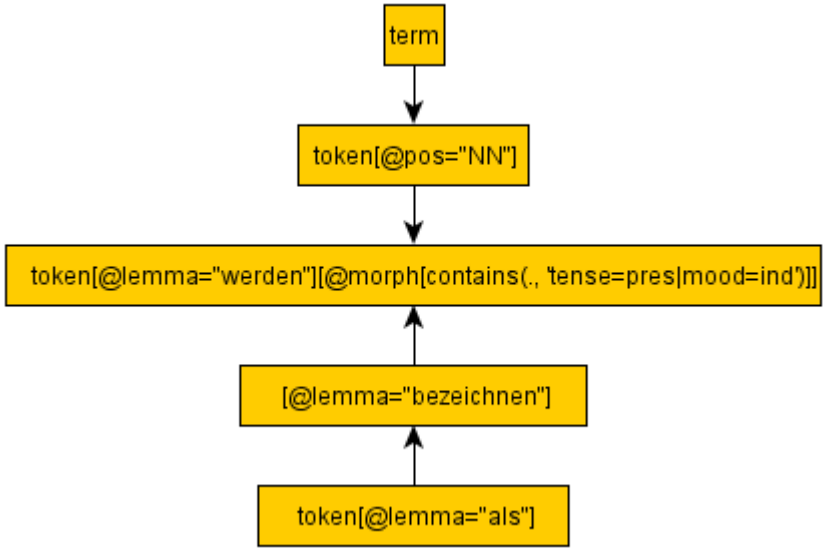
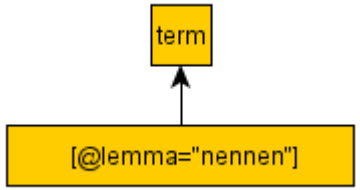
	**bestehen1** – Grundform des Musters „bestehen aus“, extrahiert z. B.: Ein <u>Wechselstromrichter</u> besteht aus elektronischen Komponenten (hauptsächlich aus IGBTs - Thyristoren und Dioden) und wandelt den Gleichstrom des Gleichrichters in die gewünschte Spannung und Frequenz um. (Sub-EKD)[1]
	**bestehen3** – diese Variante soll insbesondere leicht irreguläre Formen des Musters mit abhängigem Definiendum (Genitivattribut) identifizieren, sie extrahiert z. B.: Der Aufbau der <u>Synchrongeneratoren</u> besteht aus einem außen feststehenden Stator und innen einem Läufer (Anker), der sich auf der drehbaren Welle befindet. (Sub-EKD)
	bestehen5 – dieses Muster identifiziert Varianten mit dem Partizip I des *knowledge patterns*, z. B.: Die optimale Lösung für ein autarkes System ist eine <u>Hybridanlage</u>, bestehend aus PV- und Windgenerator, die beide einen großen Solarakku laden. (Sub-EKD)

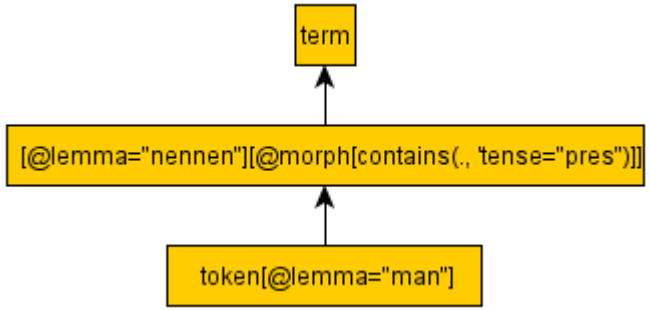
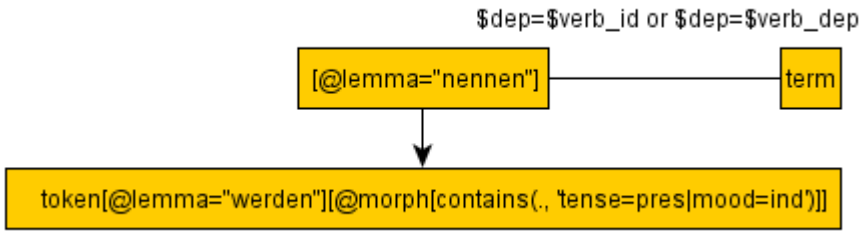
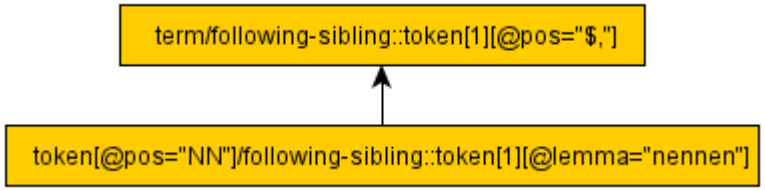
[1] Die Definienda gemäß Goldstandard werden hier und in den folgenden Beispielen durch Unterstreichung hervorgehoben, darüber hinaus können die Beispiele noch weitere Termini enthalten. Die Beispiele werden in der im Goldstandard vermerkten Form angegeben, Korpusextraktionen können dagegen zusätzlichen *noise* enthalten.

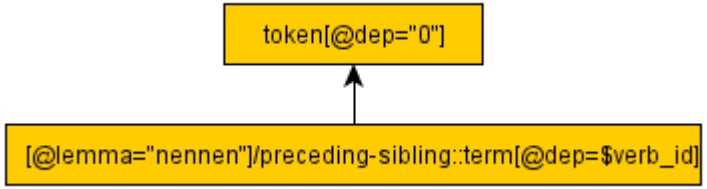
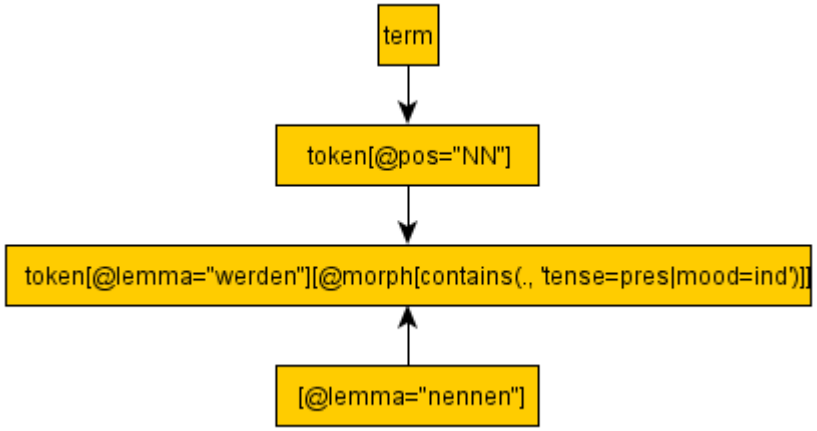
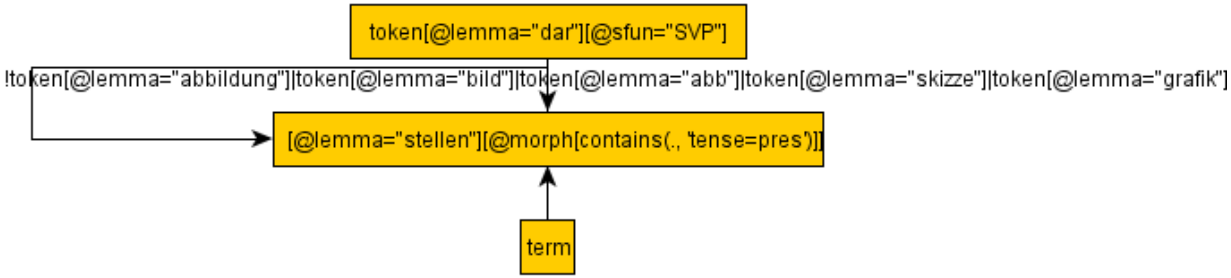
	**bestehen6** – dieses Muster soll insbesondere Varianten mit dem Modalverb „können“ identifizieren, z. B.: <u>Offshore-Fundamente</u> können aus Beton, sowie aus Stahlpfeilern oder Mehrbeinstrukturen bestehen. (Sub-EKD)
	**gegeben_sein** – hierbei handelt es sich um eine seltene Variante der Muster auf „sein“ mit der Formulierung „gegeben sein“, z. B.: Der maximale <u>Wirkungsgrad</u> eines Windgenerators ist durch das Betzsche Gesetz gegeben, das den Wirkungsgrad auf 16 / 27 » 0,59 begrenzt. (sic!) (Sub-EKD)
	**sein1a** – dies ist die Grundkonstellation der Muster auf „sein“ mit einem Definiendum und einem Subjekt, z. B.: <u>Biomasse</u> ist zusammen mit der Wasserkraft weltweit die wichtigste erneuerbare Energiequelle. (Sub-EKD) Die Drehzahl, bei der der Motor im Nennbetrieb arbeitet, ist die <u>Nenndrehzahl</u> nN. (Sub-TKD)

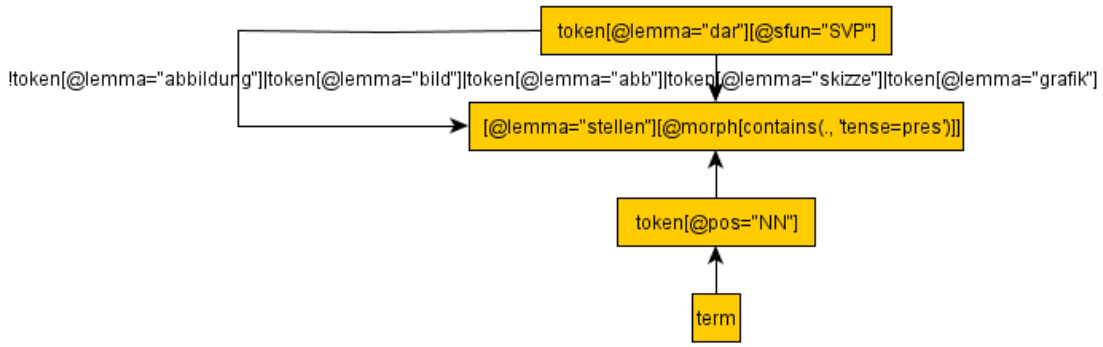
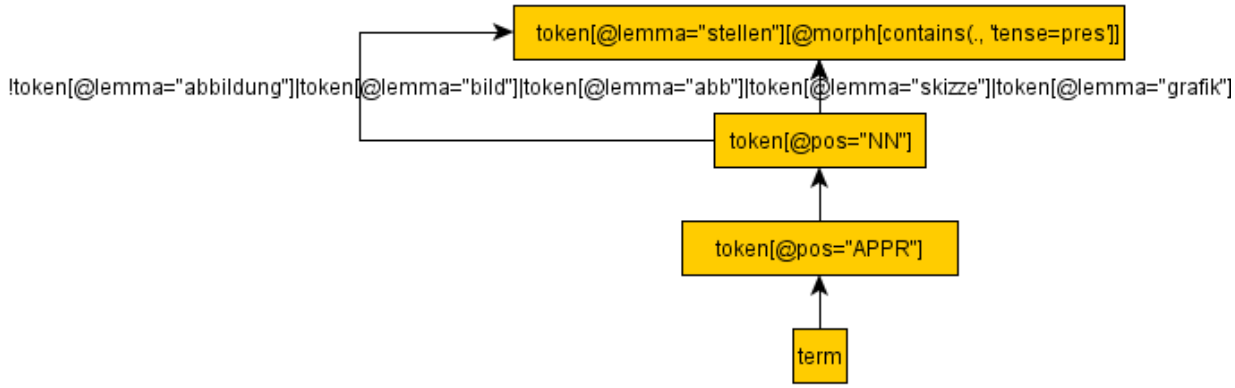
	**sein1b** – eine Variante desselben Musters, die inkonsistente Annotationen des *parsers* in Bezug auf Subjekt und Prädikatsnomen abfangen soll, z. B.: Die Brennstoffzelle ist ein Gleichstromgenerator, der Strom auf der Basis von Methanol erzeugt. (Sub-EKD)
	**bezeichnen1** – Grundform des Musters „bezeichnen“, z. B.: Der theoretische Wirkungsgrad bezeichnet den theoretisch maximalen Wirkungsgrad, den eine Energiemaschine erreichen kann. (Sub-EKD)
	**bezeichnen2** – eine weitere häufige Form des Musters mit dem Indefinitpronomen „man“, z. B.: Mit Ultraschall (US) bezeichnet man Schall mit Frequenzen, die oberhalb des vom Menschen wahrgenommenen Bereiches liegen. (Sub-EKD) Den Übergang vom flüssigen in den gasförmigen Aggregatzustand bezeichnet man als Verdampfung (oberhalb des Siedepunktes) oder Verdunstung (unterhalb des Siedepunktes) , den umgekehrten Übergang vom gasförmigen in den flüssigen Aggregatzustand als Kondensation . (sic!) (Sub-TKD)

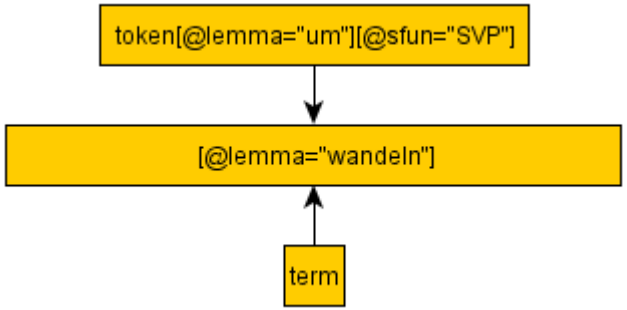
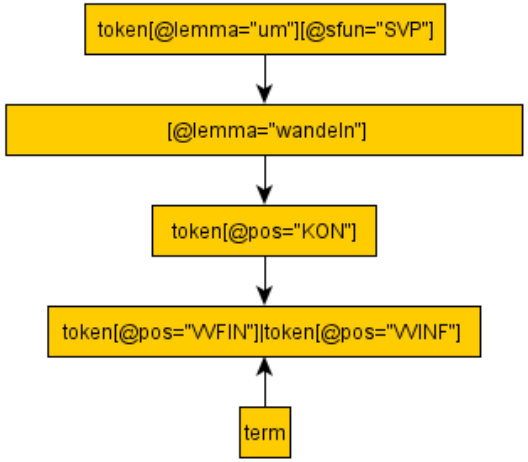
	**bezeichnen3** – die ebenfalls häufige passivierte Variante, z. B.: Als Offshorebauwerk werden feststehende Bauwerke bezeichnet, die in der offenen See vor der Küste errichtet wurden (beispielsweise Bohrinseln und Windenergieanlagen). (Sub-EKD) Ein gerichteter Elektronenstrom wird als Elektronenstrahl bezeichnet .(sic!) (Sub-TKD)
	**bezeichnen5** – eine eher seltene Mustervariante, in der das *knowledge pattern* als zweites, koordiniertes und passiviertes Prädikat fungiert, z. B.: Der NEMP tritt bei bodennahen sowie atmosphärischen Kernwaffendetonationen auf und wird entsprechend als ENDO - Nemp oder als EXO-NEMP bezeichnet. (sic!) (Sub-EKD)

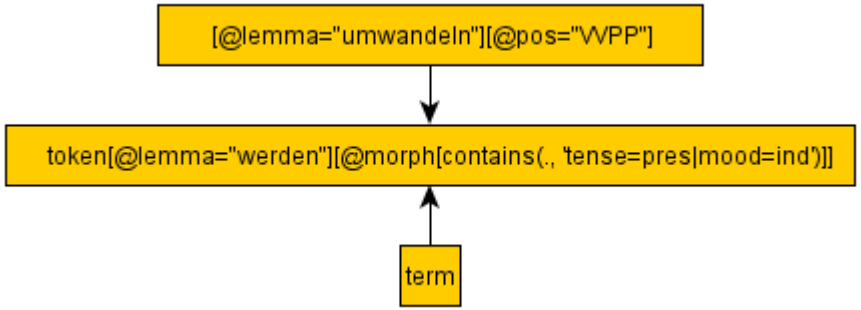
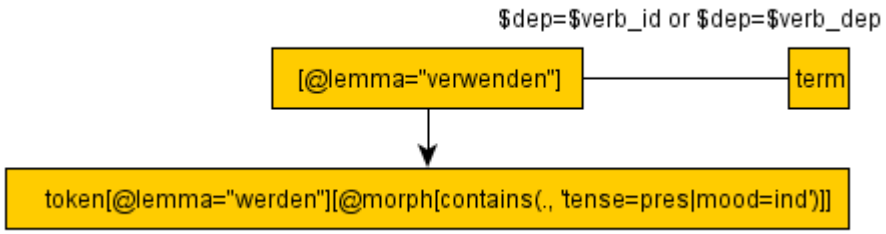
	**bezeichnen7** – eine Musterkonstellation mit abhängigem Terminus und Passivierung des *knowledge patterns*, z. B.: Die Nutzung der Windenergie im Meer wird als Offshore-Windenergie bezeichnet. (Sub-EKD) Die Einstufung der Priorität einzelner Aufgaben wird als Prioritätensetzung oder Priorisierung bezeichnet und ist wesentlicher Bestandteil des Zeitmanagements . (sic!) (Sub-TKD)
	**nennen1** – eine sehr einfache Form der Muster auf „nennen“, z. B.: Nach dem Bundesimmissionsschutzgesetz darf der Schattenwurf (auch Schlagschatten genannt) durch Windkraftanlagen auf (bestehende) Wohnhäuser jeweils nicht mehr als 30 Stunden pro Jahr und 30 Minuten pro Tag betragen. (Sub-EKD)

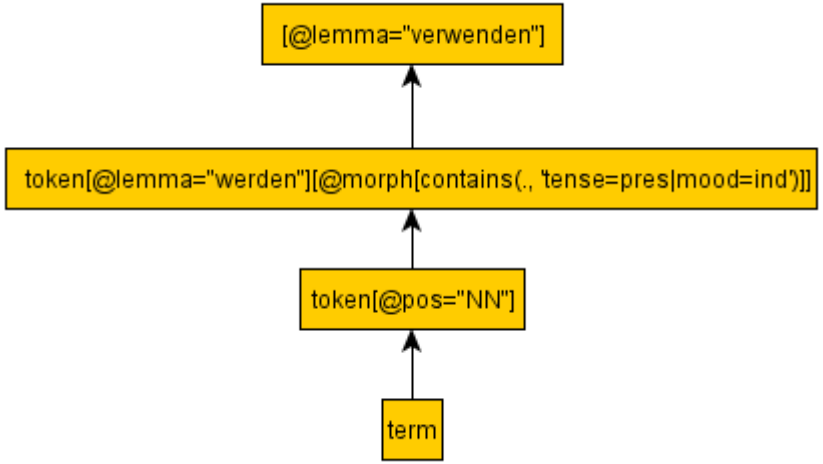
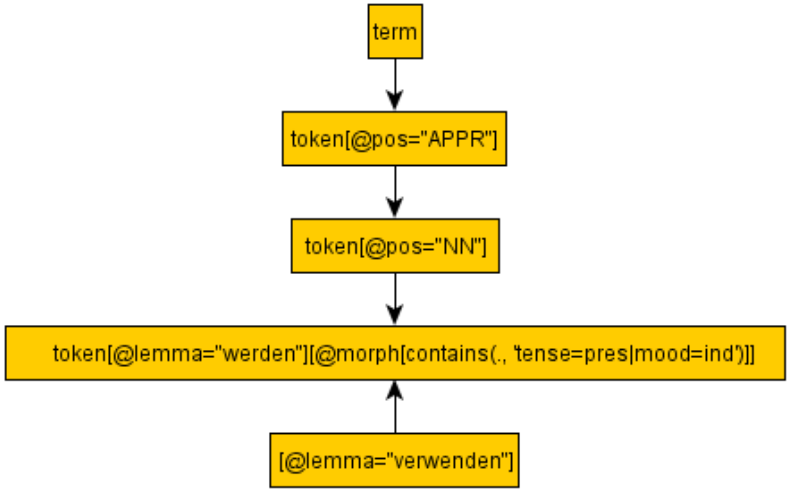
	**nennen2** – die Form des Musters auf „man“, z. B.: Die modernen drei-, zwei- oder einblättrigen Windrotoren haben hohe Schnelllaufzahlen, man nennt sie Schnellläufer. (Sub-EKD) Das Verhältnis von Schlupfdrehzahl zu synchroner Drehzahl nennt man Schlupf. (Sub-TKD)
	**nennen3** – die passivierte Form des Musters, z. B.: Infraschallereignisse im Zusammenhang mit Wettererscheinungen und Seegang werden Mikrobarome genannt. (Sub-EKD) Das einfachste Wasserstoff - Isotop 1 H besitzt keine Neutronen im Kern und wird gelegentlich Protium genannt . (sic!) (Sub-TKD)
	**nennen4** – das Muster als Apposition, z. B.: Asynchrongeneratoren, auch Induktionsgeneratoren genannt, sind ursprünglich als Elektromotoren konzipiert worden und werden dem Synchrongenerator aufgrund günstigerer Kosten und der Robustheit vorgezogen. (Sub-EKD)

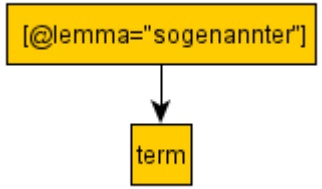
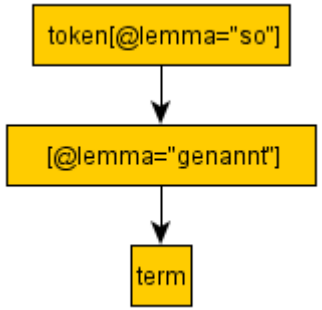
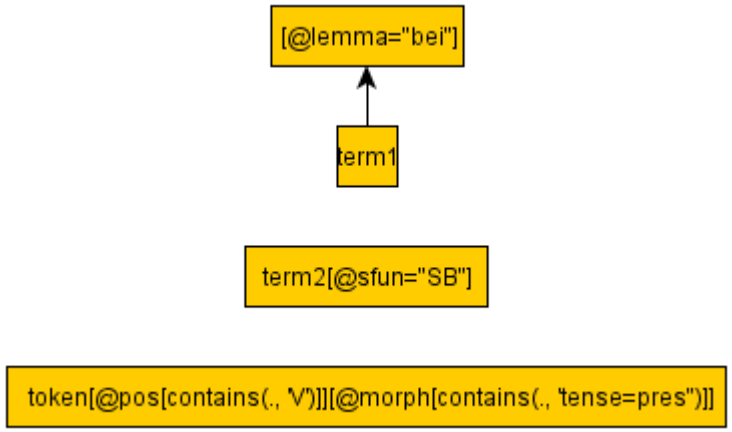
	**nennen5** – eine eher seltene Sonderform des Musters, z. B.: Aufwendiger in der Ausführung ist der von Just untersuchte Rotor mit zyklisch anstellwinkelgesteuerten Rotorblättern, auch Giromill genannt. (Sub-EKD)
	**nennen6** – eine Variante des Musters für abhängige Definienda, die aber auch Mehrwort-Definienda erkennt, z. B.: Der Wert, der die dem Wind entnommene Leistung ins Verhältnis mit der im Wind enthaltenen Leistung setzt, wird Betzscher Leistungsbeiwert () genannt und wurde von Albert Betz im Jahre 1926 ermittelt (siehe: Betzsches Gesetz). (sic!) (Sub-EKD)
	**darstellen1** – die Grundform der Muster auf „darstellen“, die versucht, Verweise auf *bildliche* Darstellungen auszuschließen, z. B.: Eine für größere Wassertiefen geeignete Gründungsstruktur stellt das Tripod-Konzept dar. (Sub-EKD) Eine Sonderform der indirekten Heizung stellen indirekt - strahlungsgeheizte Kathoden dar . (sic!) (Sub-TKD)

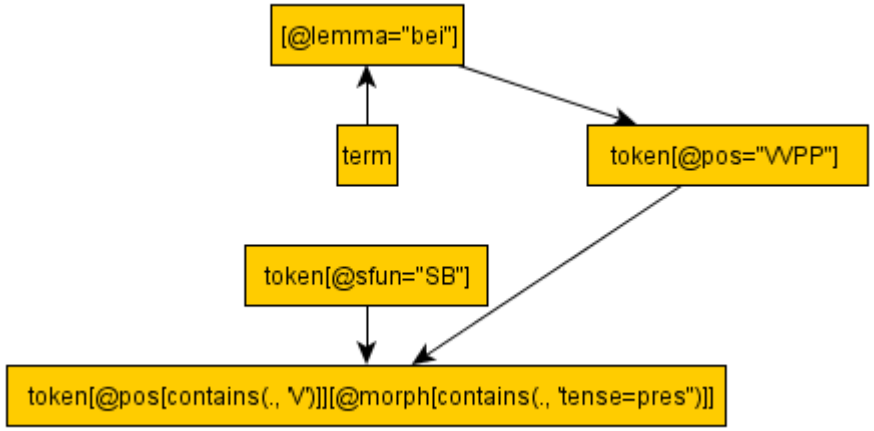
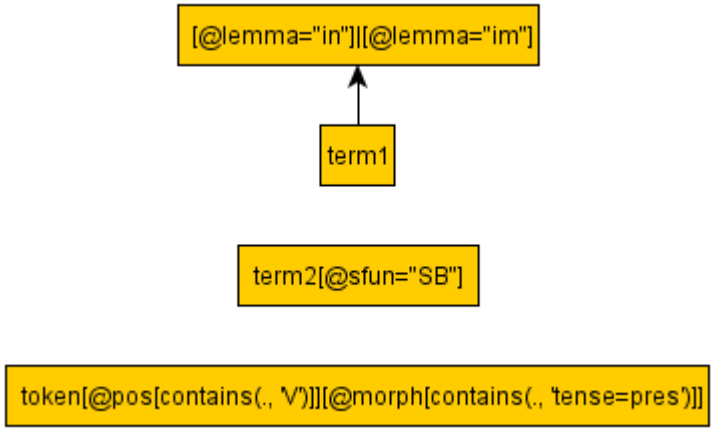
	**darstellen4** – die Variante desselben Musters mit abhängigem Definiendum, z. B.: Die Nutzung der <u>Windenergie</u> zur Stromerzeugung stellt einen vielversprechenden Ansatz dar, zukünftig einen nicht unerheblichen Anteil des weltweiten Strombedarfs aus regenerativen Quellen zu decken. (Sub-EKD)
	**darstellen5** – eine irreguläre Variante des Musters, z. B.: Der <u>Wirkungsgrad</u> stellt das Verhältnis zwischen der zugeführten Energie (hier kinetische Energie/Windenergie) zum Energieertrag (hier: elektrische Energie). (sic!) (Sub-EKD)

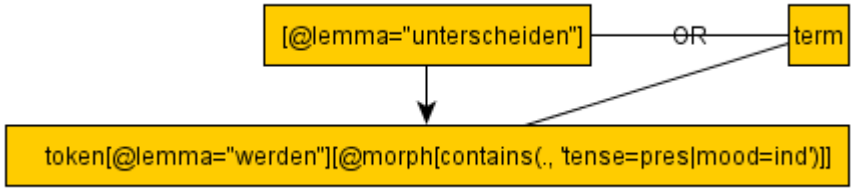
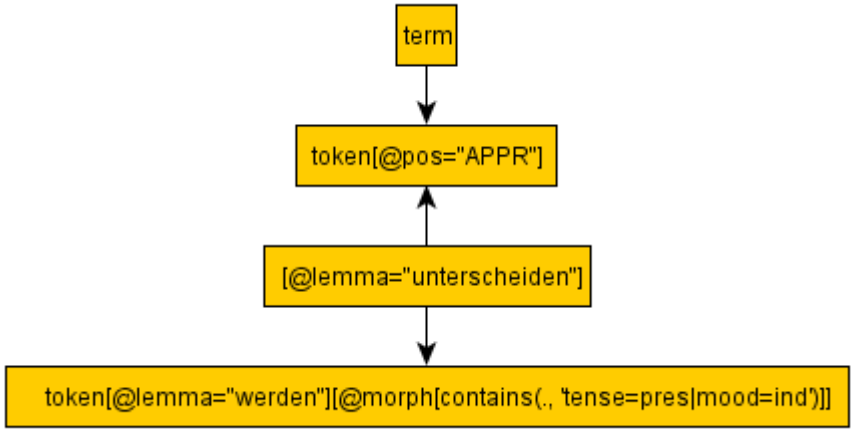
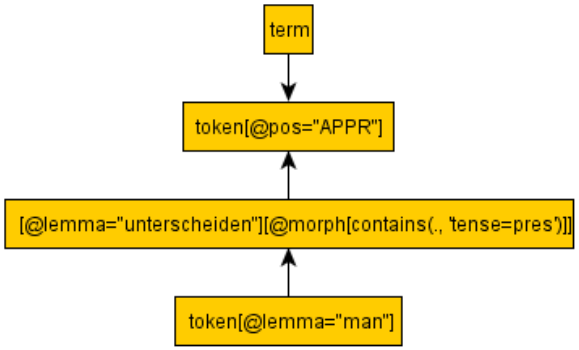
	**umwandeln1** – die Grundform des wohl fachgebietsspezifischen Musters „umwandeln“, z. B.: Der <u>Generator</u> wandelt die mechanische Drehbewegung des Triebstranges in elektrische Energie um. (Sub-EKD)
	**umwandeln2** – dasselbe Muster als koordiniertes, sekundäres Prädikat, z. B.: Der <u>Rotor</u> entzieht der strömenden Luftmasse translative, kinetische Energie und wandelt diese in mechanische Rotationsenergie um, welche an die Rotorwelle in der Triebstrangkette übertragen wird. (Sub-EKD)

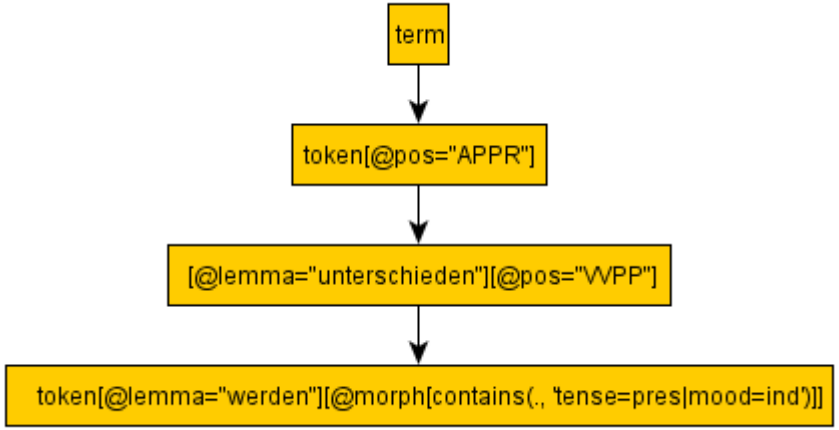
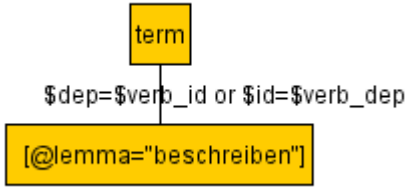
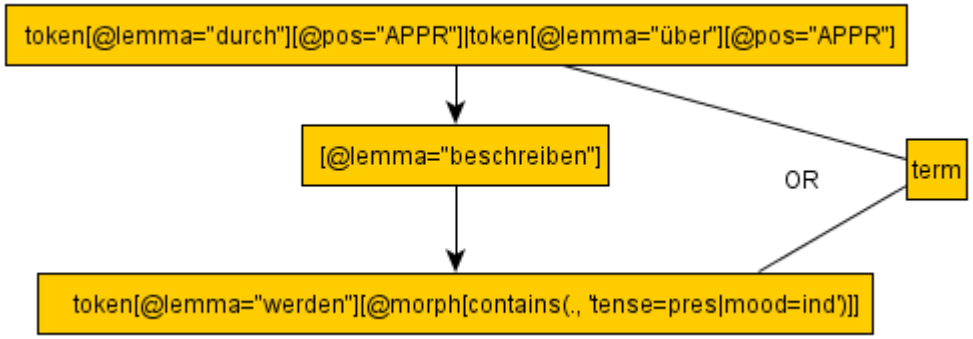
	**umwandeln4** – das Muster in seiner passivierten Variante, z. B.: Über die Rotorblätter wird die Windenergie in Rotationsenergie umgewandelt und an den Generator weitergeleitet. (Sub-EKD)
	**verwenden2** – das Muster „verwenden“ in passivierter Form, z. B.: Das Magnetron wird als selbstständiger Mikrowellengenerator in der Radartechnik und als Hauptbauteil von Mikrowellenherden verwendet . (sic!) (Sub-TKD)

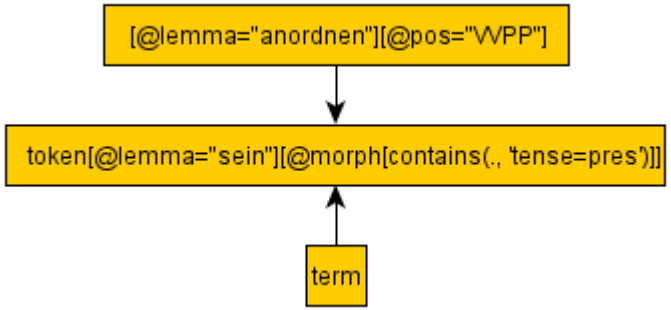
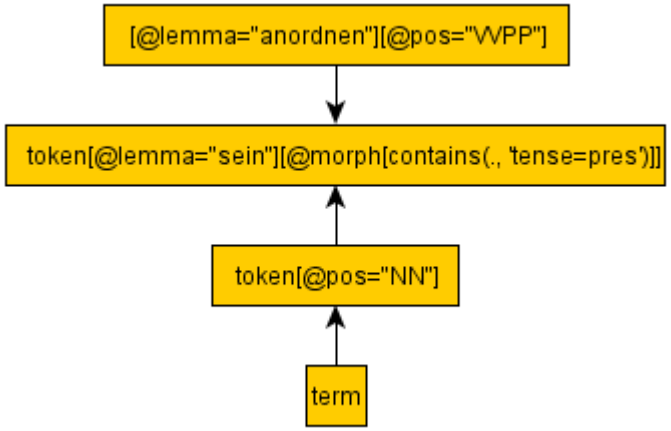
	**verwenden3** – eine Mustervariante für die Identifikation abhängiger Definienda, z. B.: Es werden prinzipiell drei verschiedene Arten von <u>Generatoren</u> zur Stromerzeugung verwendet: der Asynchrongenerator, der Synchrongenerator und der Gleichstromgenerator. (Sub-EKD)
	**verwenden4** – dasselbe Muster wiederum mit abhängigem Definiendum mit wahrscheinlich redundanten Extraktionsergebnissen

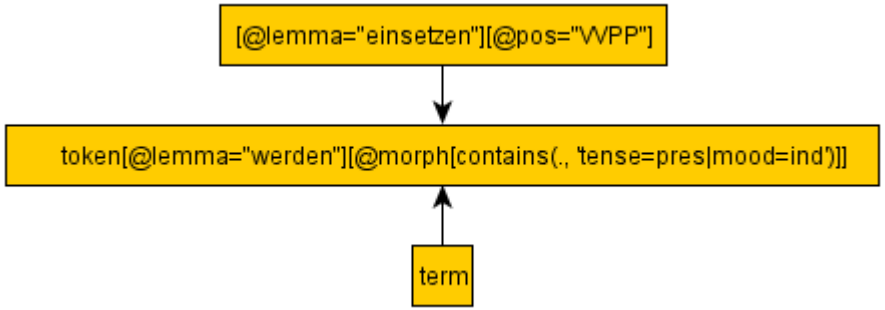
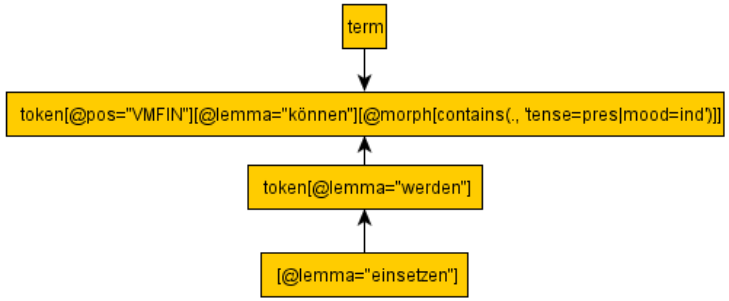
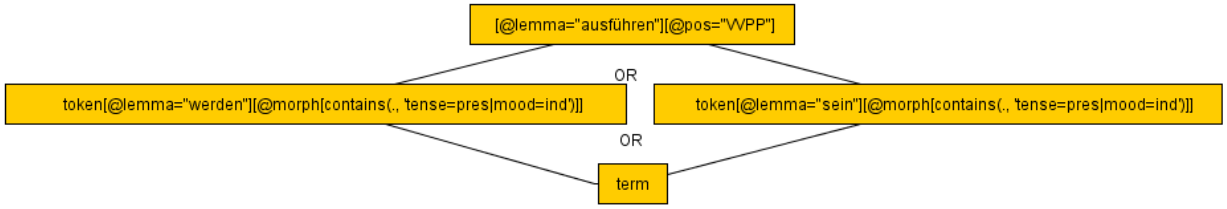
	**sogenannt1** – ein einfaches Muster zur Suche nach Termini, die durch den Marker „sogenannt“ eingeführt werden, z. B.: Bei der sogenannten Humifizierung werden stickstoffhaltige Bestandteile der Biomasse durch Mikroorganismen so abgebaut , dass unter anderem Ammoniak entsteht . (sic!) (Sub-TKD)
	**sogenannt2** – eine Variante für die Erfassung alternativer Schreibungen, z. B.: Die Regelung der Rotordrehzahl erfolgt entweder über den so genannten Stalleffekt (Strömungsabriss), oder über eine Veränderung des Anstellwinkels des Rotorblattprofils (Pitchen; von Englisch to pitch = neigen). (sic!) (Sub-EKD) Eine wichtige Eigenschaft des Wasserstoffs ist die so genannte Wasserstoffbrückenbindung , eine anziehende elektrostatische Kraft zwischen zwei Molekülen . (sic!) (Sub-TKD)
	**bei1** – das Muster „bei + Definiendum“, das allerdings in einen Satz mit Verb und terminologischem Subjekt eingebunden sein muss, z. B.: Bei halbstarren Rotorsystemen sind zwei Rotorblätter fest über ein Mittelstück verbunden, das freischwingend am Rotorkopf aufgehängt ist. (Sub-EKD)

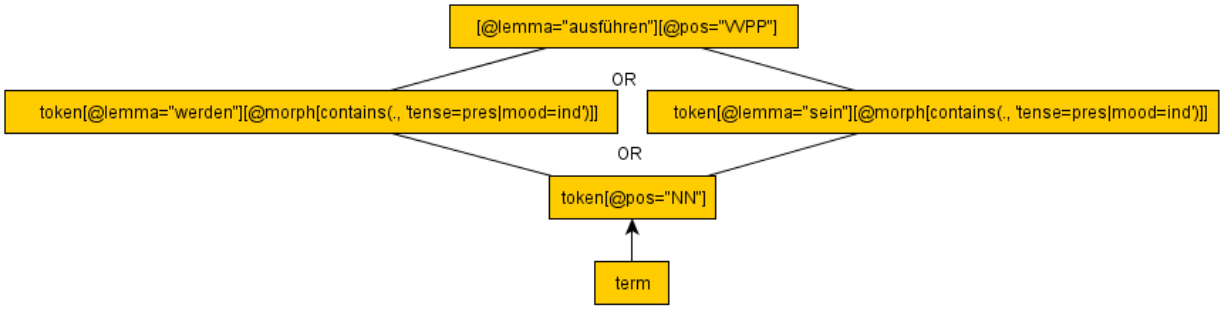
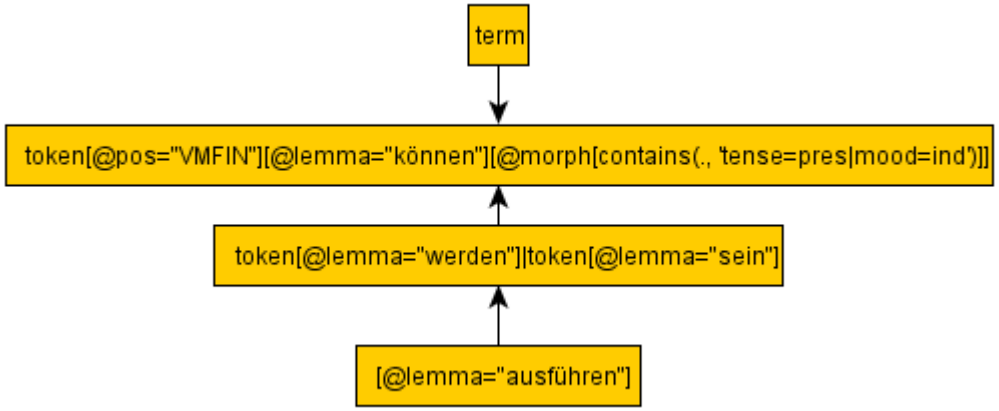
	**bei2** – eine komplexere Variante desselben Musters, z. B.: Bei modernen Windkraftanlagen zur Stromerzeugung sind die Rotorflügel aerodynamisch ausgelegt, das heißt, sie haben ein aerodynamisches Profil, das einen Auftrieb erzeugt, der in ein Drehmoment umgesetzt wird, das zum Antrieb des Generators dient. (Sub-EKD) Bei idealen Gasen ist die freie Beweglichkeit der einzelnen Teilchen entsprechend der kinetischen Gastheorie vollkommen ; dieser Zustand wird erst bei hohen Temperaturen gegenüber dem Siedepunkt erreicht (was z . B . für Wasserstoff und Helium bereits bei Zimmertemperatur gilt) . (sic!) (Sub-TKD)
	**in** – ein einfaches Muster zur Suche nach Strukturen „in + Term“ innerhalb eines vollständigen Satzes mit zweitem Terminus, z. B.: Die Bremse in der Gondel verhindert bei zu großen Windgeschwindigkeiten, dass die Anlage durch zu hohe Drehzahlen geschädigt wird. (Sub-EKD)

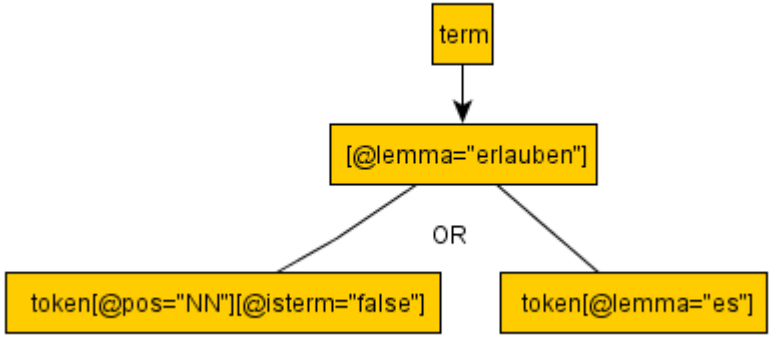
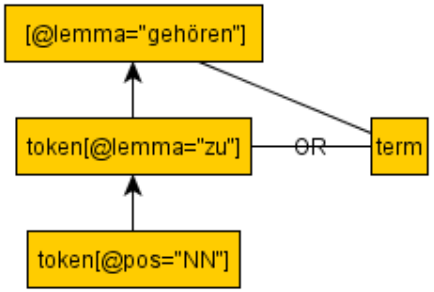
	**unterscheiden1** – die Grundform des Musters „unterscheiden“, z. B.: Die verschiedenen Röhrentypen werden durch ihre Funktion und die Anzahl und Anordnung der Elektroden unterschieden , außerdem durch unterschiedliche Stromversorgung (Netzröhren , Batterieröhren und Niederspannungsröhren) . (sic!) (Sub-TKD)
	**unterscheiden2** – eine Variante desselben Musters mit anderer Einbettung des Definiendums, z. B.: Bei der Windenergienutzung wird zwischen Windgeneratoren mit horizontaler (Horizontalachsenkonverter) und vertikaler (Vertikalachsenkonverter) Drehachse unterschieden, wobei letztere bisher keine Rolle im kommerziellen Sektor spielen. (Sub-EKD)
	**unterscheiden4** – eine ähnliche Form des Musters mit dem Indefinitpronomen „man“, z. B.: Hinsichtlich der Bauformen unterscheidet man zwischen Anlagen mit horizontaler und vertikaler Achse. (Sub-EKD)

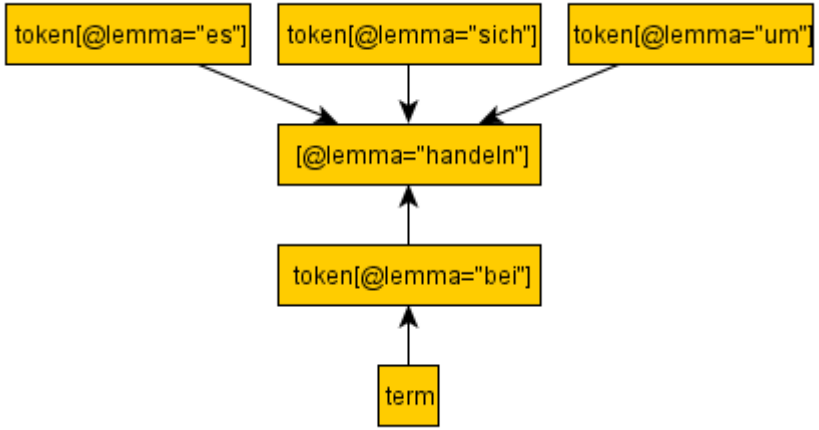
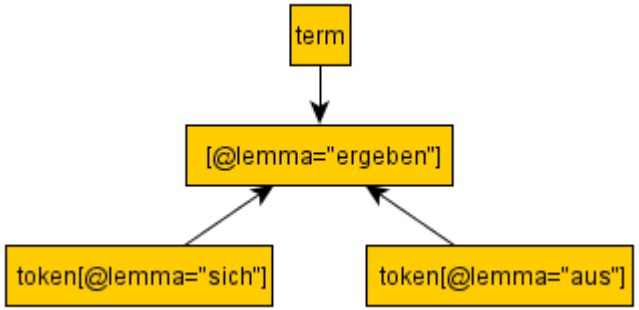
	**unterscheiden5** – die passivierte Form des Musters, z. B.: Bei der Anbindung einer Windkraftanlage an das Netz der öffentlichen Versorgung oder ein beliebiges Inselnetz wird unterschieden zwischen einer direkten und indirekten Netzkopplung. (Sub-EKD)
	**beschreiben1** – die Basis-Variante des Musters „beschreiben“, z. B.: Der sogenannte Abstandsfaktor kA beschreibt den zwischen zwei benachbarten Konvertern minimal einzuhaltenden Abstand. (Sub-EKD)
	**beschreiben2** – eine passivierte Variante desselben Musters mit anderer Einbindung des Definiendums, z. B.: Die Weibullverteilung aus der Gruppe der Gammaverteilungen wird durch den scale-Parameter c (Skalierung) und den shape-Parameter k (Form) beschrieben. (Sub-EKD) Das Verhalten von Synchronmotoren am Netz wird oft statt durch die Regelkennlinien durch die wegen ihrer Form sogenannten V-Kurven beschrieben. (Sub-TKD)

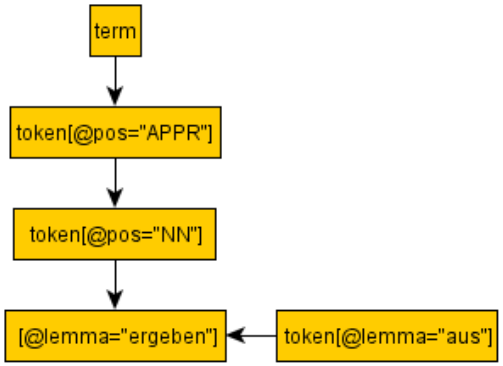
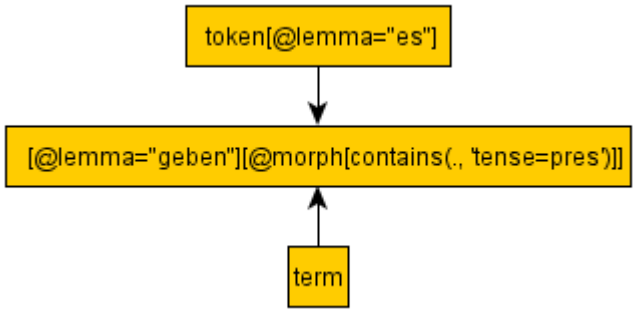
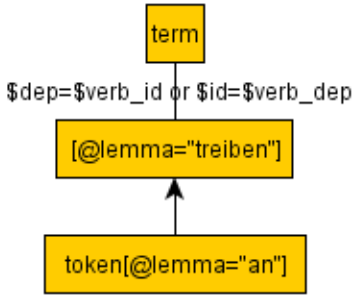
	**anordnen1** – die Grundform des Musters „anordnen“, z. B.: Die vertikalen Einleitflächenelemente sind radial angeordnet, wodurch die Windkraftanlage windrichtungsunabhängig ist. (Sub-EKD)
	**anordnen3** – dasselbe Muster mit abhängigem Definiendum, z. B.: Die zylindrische Säule des Turmes ist auf dem Dach des Maschinenraumes angeordnet. (Sub-EKD)

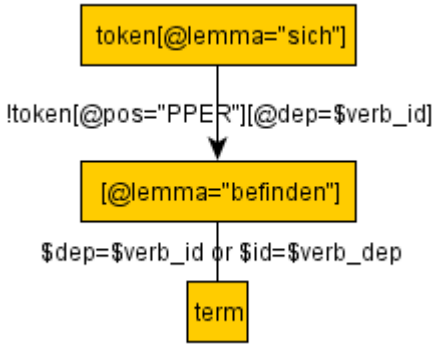
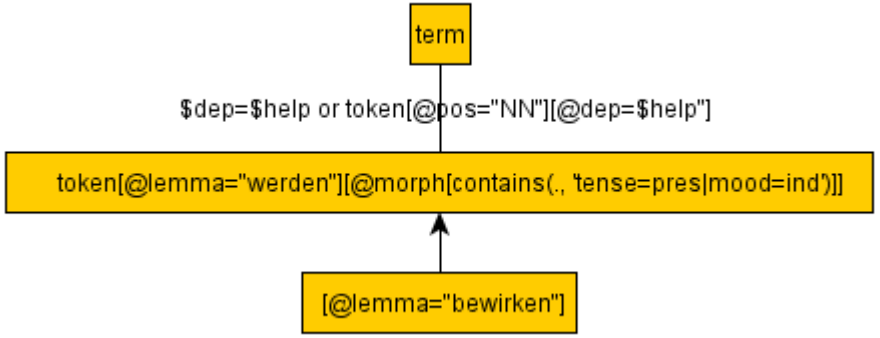
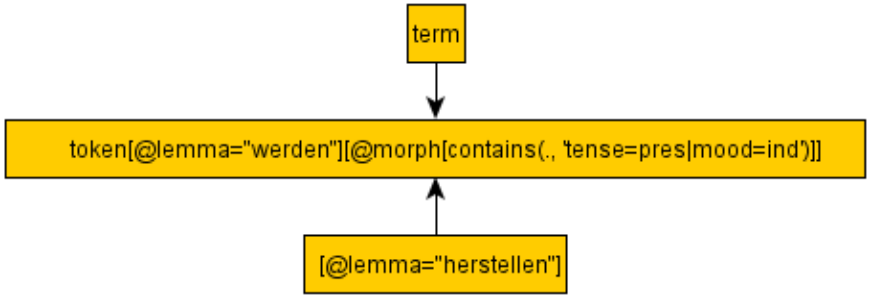
	**einsetzen1** – die Grundform des Musters „einsetzen“, z. B.: Insbesondere bei großen Windkraftanlagen wird aufgrund der diskutierten Nachteile der Stall-Regelung die Verstellung des Blattwinkels (Pitchregelung) zur Regelung der dem Wind entnommenen Leistung eingesetzt. (Sub-EKD) Direkt geheizte Kathoden werden noch heute bei Senderöhren , Gleichrichterröhren und Magnetrons eingesetzt . (sic!) (Sub-TKD)
	**einsetzen3** – die Variante des Musters mit dem Modalverb „können“, z. B.: Windenergieanlagen können in allen Klimazonen, auf See und in allen Landformen (Küste, Binnenland, Gebirge) zur Gewinnung elektrischen Stroms eingesetzt werden. (Sub-EKD)
	**ausführen1** – die Grundform des Musters „ausführen“ nutzt entweder „sein“ oder „werden“, um das Definiendum mit dem *knowledge pattern* zu verbinden, z. B.: Das Bucket-Fundament wird als eimerartiges Stahlfundament ausgeführt, das mit der Öffnung nach unten durch ein Vakuumverfahren in den Meeresboden eingebracht wird. (Sub-EKD)

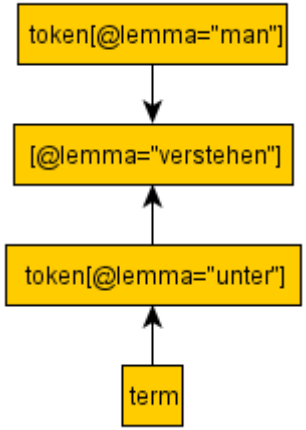
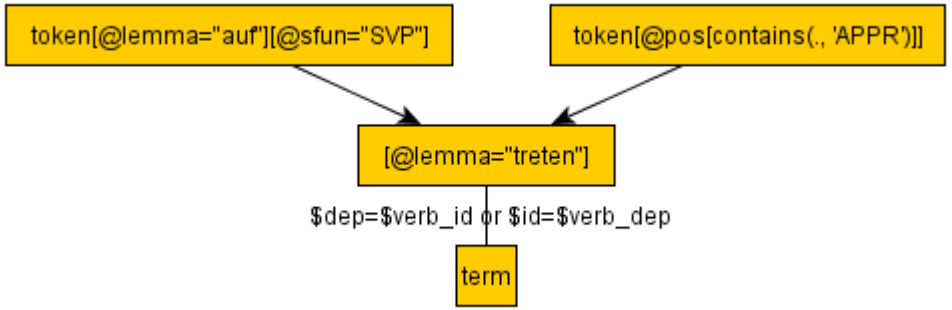
	**ausführen3** – diese Form des Musters soll abhängige Definienda identifizieren, extrahiert aber auch andere Strukturen, vgl.: Die vertikalen <u>Einleitflächenelemente</u> 12 sind in gerader Form, vorzugsweise leicht gekrümmter Form, oder auch gekrümmter Form ausgeführt, wie sie Figur 3 zeigt, wobei sie eine aerodynamische Ausgestaltung nach Figur 7 erfahren. (Sub-EKD)
	**ausführen4** – Mustervariante mit modalem „können“, z. B.: Die vertikalen <u>Einleitflächenelemente</u> können wahlweise geradförmig, als auch gekrümmt ausgeführt sein, wobei die Krümmung nicht größer sein darf als die Schrägstellung und Krümmung der Rotorflügel. (Sub-EKD)

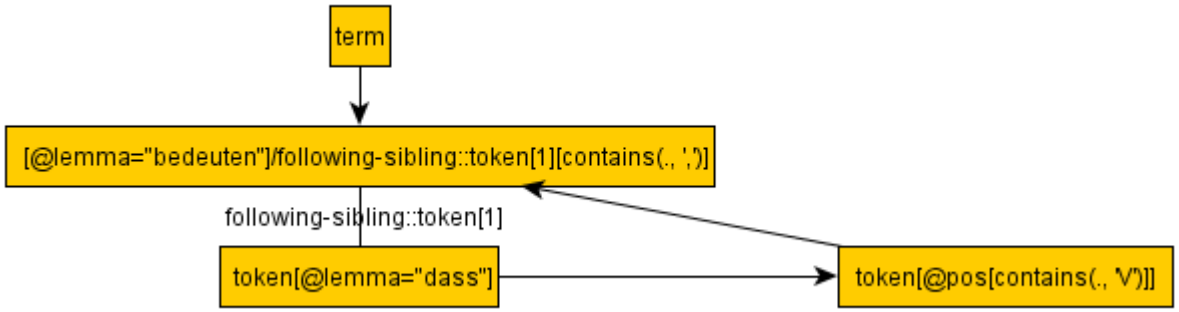
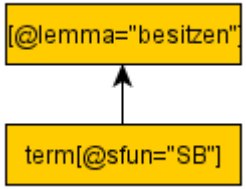
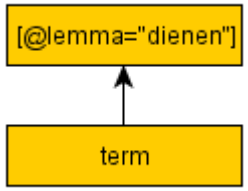
	**erlauben** – das Muster „erlauben“ erfasst sowohl nachgeschaltete Teilsätze als auch Nominalphrasen, z. B.: Die so genannte Pitch-Regelung (Blattwinkelverstellung) erlaubt es, die Rotorblätter einzeln aus dem Wind zu drehen. (Sub-EKD) Gemeinsame Normen und Standards erlauben den freien Handel von Waren und Dienstleistungen ohne zusätzliche Anpassungskosten. (Sub-TKD)
	**gehören1** – in dieser Version des Musters kann das Definiendum sowohl vom Verb als auch von der Präposition abhängen, z. B.: Die Windenergie gehört zu den erneuerbaren Energien, denn Wind steht zwar nicht stetig, aber dennoch dauerhaft zur Verfügung und wird vor allem zur Stromerzeugung genutzt.(Sub-EKD) Zu den Softkillsystemen gehören die sogenannten Jammer und Decoys. (Sub-TKD)

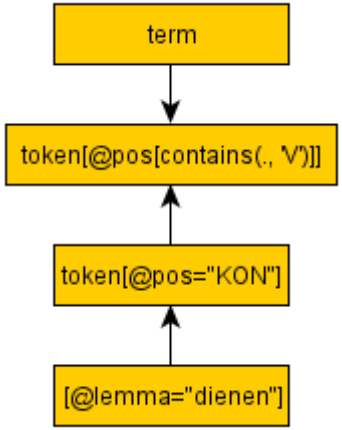
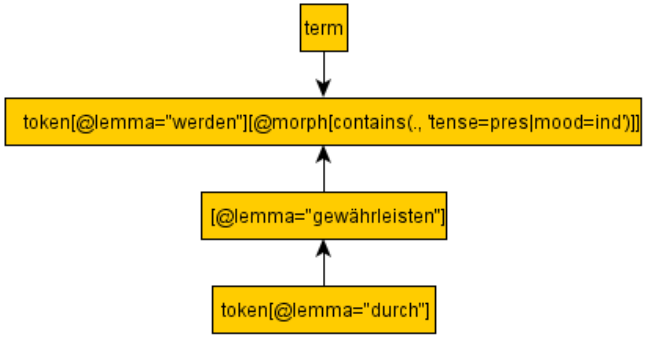
	**handeln_um** – das einfache Muster „bei + Definiendum handelt es sich um“, z. B.: Bei der Windkraft handelt es sich um die kinetische Energie der bewegten Luftmassen der Atmosphäre. (Sub-EKD) Bei einem Dampf im engeren Sinn handelt es sich um einen Gleichgewichtszustand zwischen flüssiger und gasförmiger Phase. (Sub-TKD)
	**ergeben_sich1** – ein wahrscheinlich eher seltenes Muster, z. B.: Der Anstellwinkel der Flügel des Rotors ergibt sich aus dem Abstand der Rotorachsenmitte zum jeweiligen Flügelende und der vorderen Begrenzung der Flügelnase die an die Linie stößt, welche durch die Mitte der Rotorachse läuft. (sic!) (Sub-EKD)

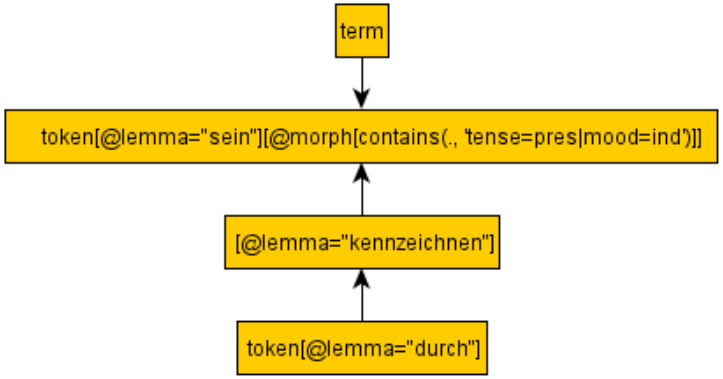
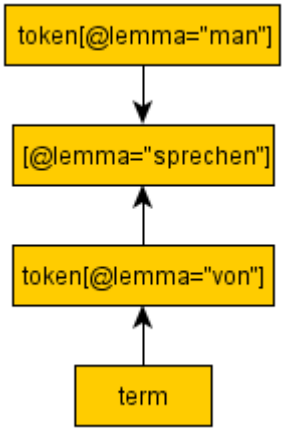
	**ergeben_sich2** – die Variante desselben Musters für die Erkennung abhängiger Definienda, z. B.: Der Bedarf an windenergiebedingter Regelleistung ergibt sich aus der Differenz der prognostizierten Einspeisung aus Windenergieanlagen und den tatsächlichen Einspeisewerten. (Sub-EKD)
	**es_gibt** – das Muster zur Beschreibung thetischer Präsentative, z. B.: Weltweit gibt es viele verschiedene Winde und Windsysteme, wie zum Beispiel den Passat, Monsun, Föhn, den Mistral, die Bora oder den Scirocco. (Sub-EKD) Die Photozelle gibt es in Vakuumausführung oder gasgefüllt , sie ändert in Abhängigkeit vom auf sie einfallenden Licht ihren Elektronenstrom (Photoeffekt) . (sic!) (Sub-TKD)
	**antreiben1** – die Grundform des Musters „antreiben“, z. B.: Der Rotor treibt über die gemeinsame Welle den Generator an. (Sub-EKD)

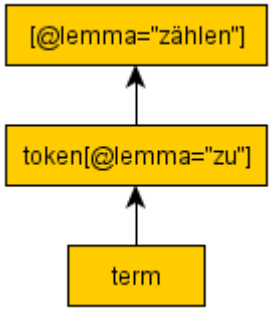
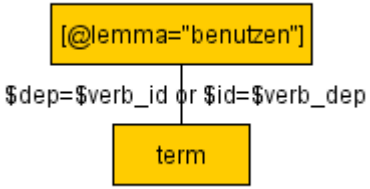
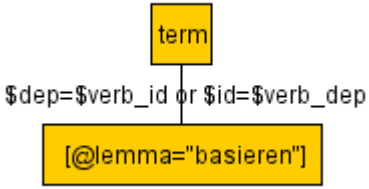
	**befinden** – das einfache Muster „befinden“, das u. U. auch unerwartete Extraktionsergebnisse produzieren kann, z. B.: Man unterscheidet, ob sich der Rotor auf der dem Wind zugewandten Seite (Luvläufer) oder auf der dem Wind abgewandten Seite (Leeläufer) des Turmes befindet. (Sub_EKD) Die Torpedoräume , in denen die Torpedos und andere Waffen gelagert werden , befinden sich meist im Bug des U - Bootes . (sic!) (Sub-TKD)
	**bewirken2** – eine Variante des Musters „bewirken“, in der das Definiendum entweder vom Hilfsverb „werden“ ($help) oder von einem auf dieses bezogenen Nomen abhängig ist, z. B.: Eine Verstärkung des Auftriebes wird durch die Erhöhung des so genannten „Anstellwinkels“ bewirkt, welcher den Winkel der Querschnittlinie (Profilsehne) des Rotorblattes/ Tragfläche zu der anströmenden Luft beschreibt. (sic!) (Sub-EKD)
	**herstellen** – das einfache Muster „herstellen“, z. B.: Die Rotorblätter werden heute überwiegend aus Faserverbundwerkstoffen hergestellt und können eine Länge bis zu 30 Metern haben. (Sub-EKD)

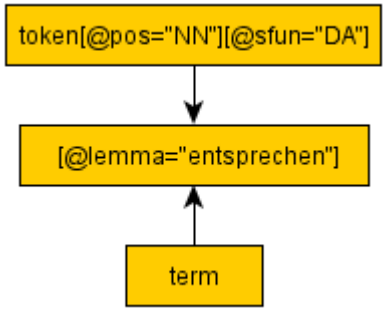
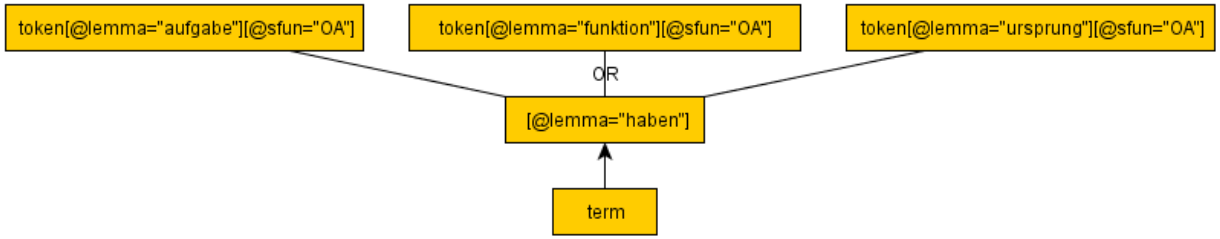
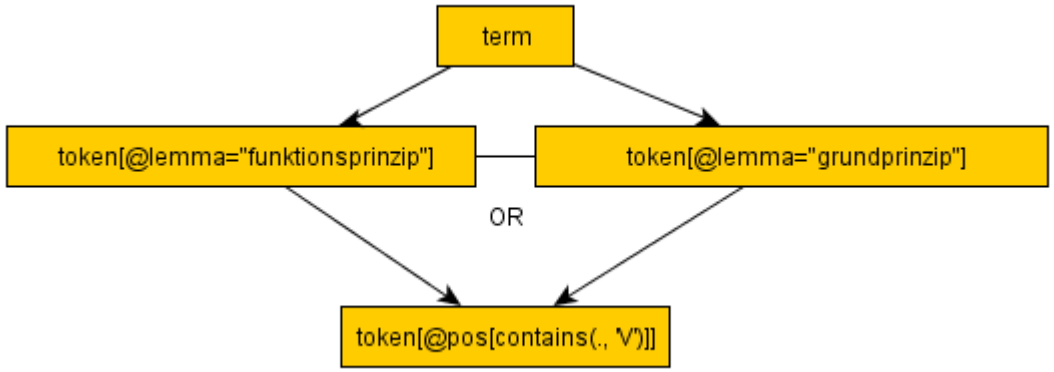
	**verstehen** – das einfache Muster „unter + Definiendum versteht man“, z. B.: Unter regenerativen oder erneuerbaren Energiequellen versteht man all diejenigen Energienutzungsschemata, die ohne die Ausbeutung nur begrenzt vorhandener Vorkommen an Primärenergiequellen auskommen. (Sub-EKD)
	**auftreten** – ein einfaches Muster für das Prädikat „auftreten“, z. B.: Infraschall, also Schall mit Frequenzen zwischen 0 Hz und 16 Hz, tritt z.B. im Zusammenhang mit Erdbeben oder tektonischen Verschiebungen als seismische Welle auf. (Sub-EKD)

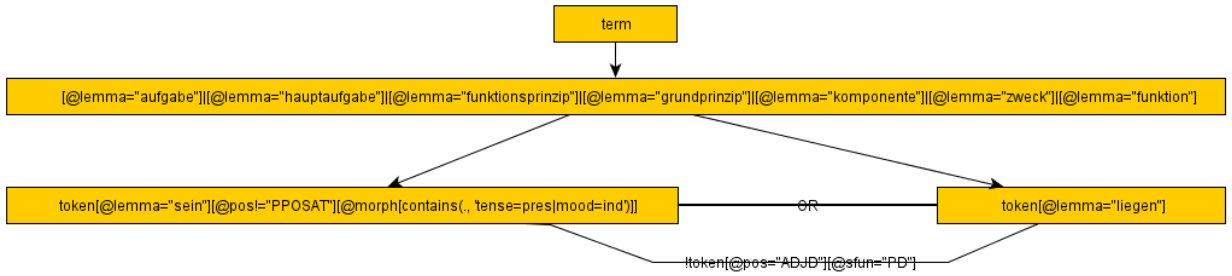
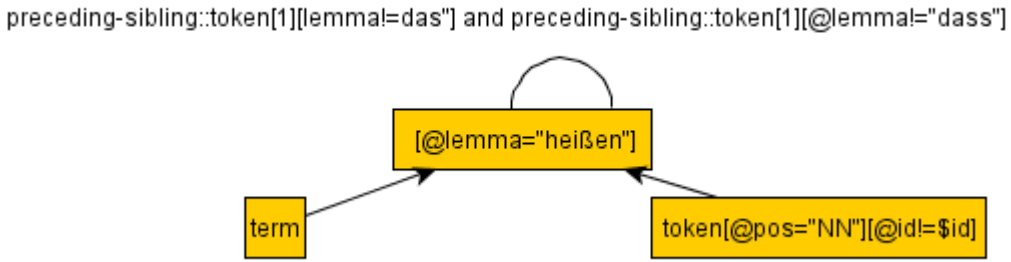
	**bedeuten** – das Muster extrahiert die Struktur „Definiendum + bedeutet, dass + V“, z. B.: Inselbetrieb bedeutet, dass keine Verbindung zum Netzt der öffentlichen Stromversorgung besteht, wie zum Beispiel bei abgelegenen Häusern in wenig besiedelten Landstrichen. (Sub-EKD)
	**besitzen** – ein einfaches Muster für meronymische Beziehungen, z. B.: Die Anlagen besitzen auch eine Notstromversorgung, um bei Netzausfall ein sicheres Abschalten (Blätter in Segelstellung drehen und/oder bremsen) zu gewährleisten. (Sub-EKD) Die einfachsten Einphasen-Induktionsmotoren besitzen im Ständer nur einen Wicklungsstrang. (Sub-TKD)
	**dienen1** – ein einfaches Muster für die Angabe funktionaler Beziehungen, z. B.: Die auch bei Windkraftanlagen mit mehr als 100 Metern Höhe vorgeschriebene Hindernisbefeuerung dient der Sicherheit des Flugverkehrs. (Sub-EKD)

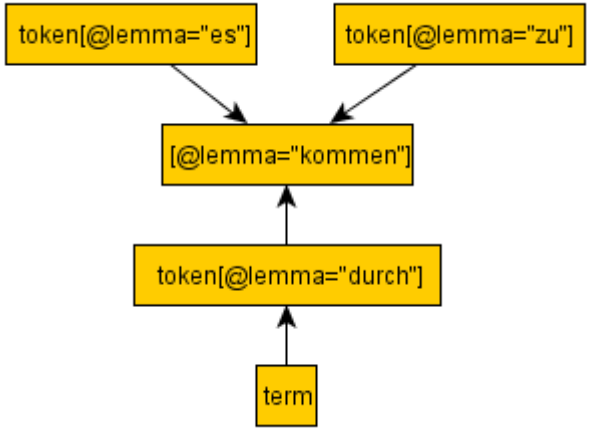
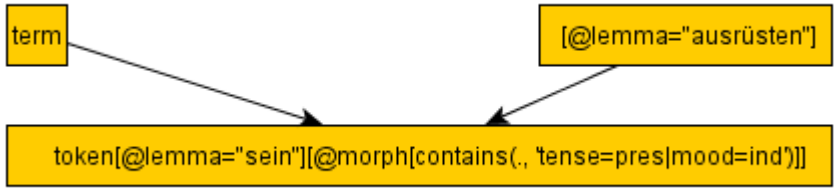
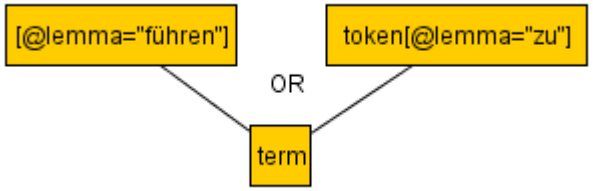
	**dienen2** – ein Muster für das koordinierte Prädikat „dienen“, z. B.: Die vorgespannten Zugstränge 6 sind den Zugsträngen 5 entgegen gerichtet angeordnet und dienen dazu, dass die Rotorblätter 4 auch bei geringen Windgeschwindigkeiten oder Windstille nicht knicken und die Gewichtskräfte der Rotorblätter abgefangen werden. (Sub-EKD)
	**gewährleisten2** – ein Muster für die passivierte Variante des Verbs „gewährleisten“, z. B.: Die Windrichtungsnachführung wird durch hydraulische Motoren oder Elektromotoren (Azimutantrieb) gewährleistet. (Sub-EKD)

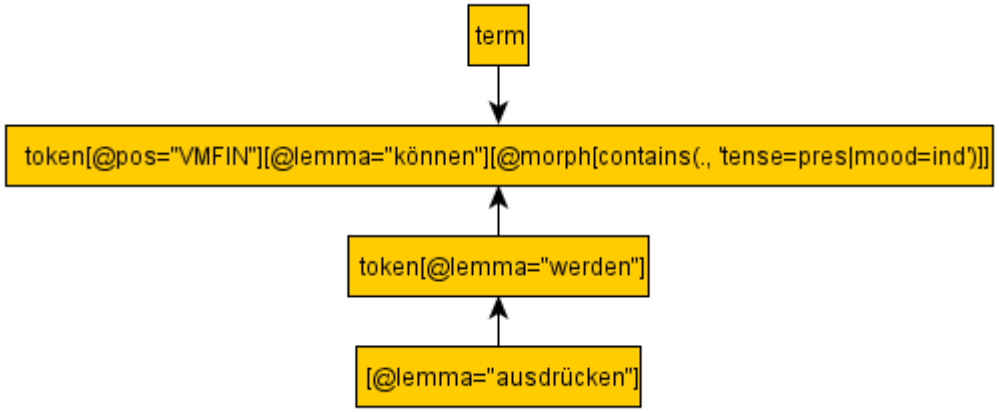
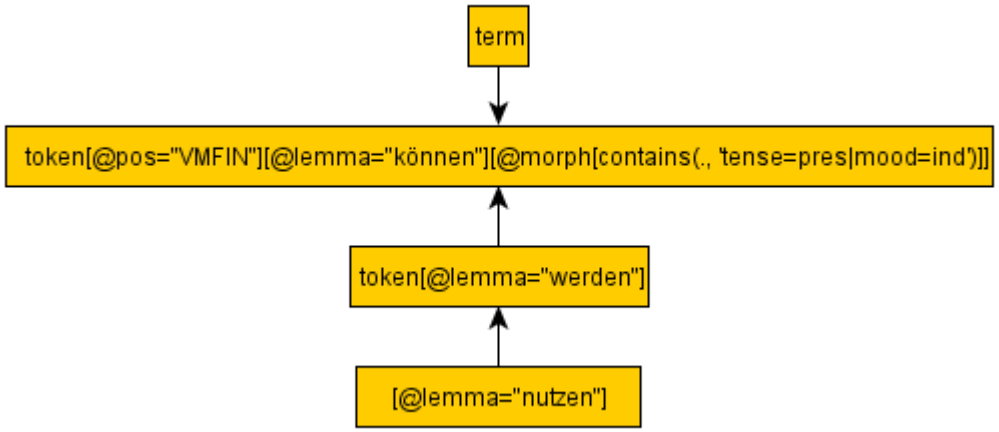
	**kennzeichnen** – ein einfaches Muster für die Abfrage von Referenz-Beziehungen, z. B.: Die Darrieus-Maschine ist durch ihre C-förmigen Rotorblätter gekennzeichnet, die an einen Schneebesen erinnern. (Sub-EKD)
	**sprechen** – dieses Muster identifiziert die Angabe von Bedingungen für die Klassenzugehörigkeit von Begriffen mit der Formulierung „man spricht von + Definiendum“, z. B.: Von Inselbetrieb spricht man, wenn keine Verbindung zum öffentlichen Stromversorgungsnetz besteht. (Sub-EKD) Wird Strom über ein räumlich begrenztes Stromnetz verteilt , spricht man von einem Inselnetz . (sic!) (Sub-TKD)

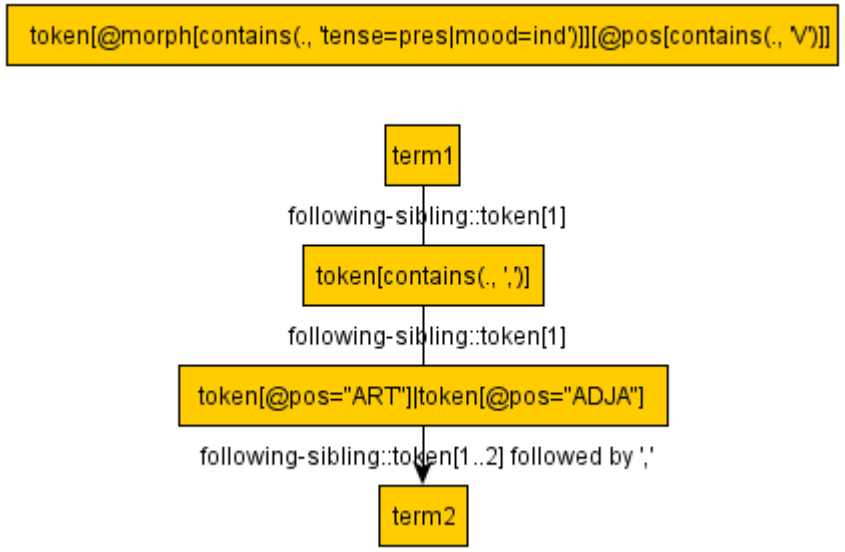
	**zählen** – ein einfaches Muster für die Angabe von Klassenzugehörigkeiten, z. B.: Zu den Langsamläufern zählen Anlagen mit geringen Drehzahlen und hoher Blattzahl; z.B. Holländer-Windmühlen oder amerikanische Westernrotoren. (Sub-EKD) Mündliche Beschreibungen und im Ausland erfolgte Benutzungshandlungen zählen somit nicht zum Stand der Technik . (sic!) (Sub-TKD)
	**benutzen** – ein einfaches und generisches Muster für unterschiedliche semantische Beziehungen, z. B.: Die Schwerkraftgründungen benutzen die Schwerkraft, um die Anlage auf dem Meeresboden zu fixieren. (Sub-EKD) Aufgrund seiner hohen Wärmekapazität benutzt man Wasserstoff in Kraftwerken und den dort eingesetzten Turbogeneratoren als Kühlmittel . (sic!) (Sub-TKD)
	**basieren** – ein einfaches Muster für das Verb „basieren“, z. B.: Das sogenannte Aufwindkraftwerk basiert auf dem Grundgedanken, eine Luftströmung wie in der Natur durch Erwärmung, das heißt Unterschiede in der Luftdichte, herbeizuführen. (Sub-EKD)

	**entsprechen** – ein ebenfalls generisches, einfaches Muster für das Verb „entsprechen“, z. B.: Die tatsächlich genutzte Windenergie PWind entspricht der Differenz der Windenergien vor und hinter dem Konverter. (Sub-EKD)
	**haben** – dieses Muster sucht bestimmte Funktionsverbgefüge mit dem Verb „haben“, z. B.: Die vertikalen Einleitflächenelemente 12 haben die besondere Aufgabe, die in der Windrichtung zurücklaufenden Rotorflügel 1, 2, 3 abzudecken und den Wind zusätzlich auf die positive Seite zu leiten. (Sub-EKD)
	**prinzip** – in diesem Muster ist der Terminus syntaktisch von einem der beiden Lemmata „Funktionsprinzip“ oder „Grundprinzip“ abhängig, die sich wiederum auf ein Verb beziehen sollten, z. B.: Das Grundprinzip des Blattspitzenantriebs ist, wie der Name schon sagt, das Rotorblatt durch eine Kraft an der Spitze anzutreiben. (Sub-EKD)

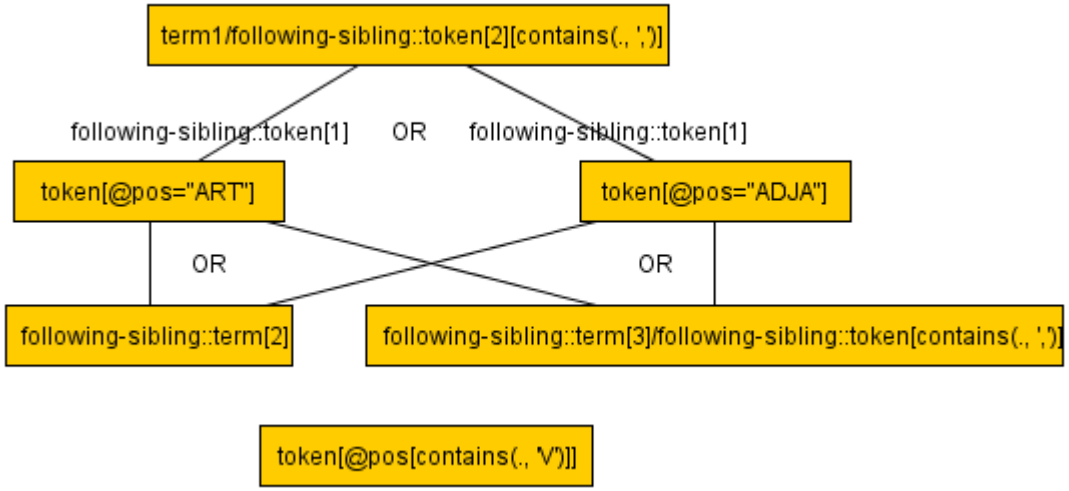
	**etwas_sein** – dieses Muster extrahiert eine Reihe unterschiedlicher Zuschreibungen, z. B.: Der Zweck des Rotors liegt natürlich darin, die lineare Bewegung des Windes in eine Drehbewegung umzusetzen, die zum Antrieb eines Generators verwendet werden kann. (Sub-EKD)
	**heißen** – dieses Muster sucht das Prädikat „heißen“ in Bezug auf das Definiendum und ein weiteres Nomen und schließt dabei die Sequenz „das heißt“ einschließlich möglicher falscher Schreibungen aus, z. B.: Der Unterschied zwischen den, von verschiedenen Abschnitten eines Rotorblatts erzeugten, Auftriebswerten heißt Auftriebs-Asymmetrie. (sic!) (Sub-EKD) Der direkte Übergang vom festen in den gasförmigen Aggregatzustand ist die Sublimation , der umgekehrte Übergang vom gasförmigen in den festen Aggregatzustand heißt Resublimation . (sic!) (Sub-TKD)

	**kommen_zu** – ein einfaches Muster für die Extraktion von Kausalbeziehungen, z. B.: Durch die Ionisation von Atomen kommt es zu Veränderungen in den Zellen, die Störungen der Zelltätigkeit und sogar einen völligen Zelltod bewirken können. (Sub-EKD)
	**ausrüsten** – ein Muster für das Prädikat „ausrüsten“, z. B.: Schnellläufer sind mit langen und leichten Blättern ausgerüstet, die aus Verbundmaterial (Glasfaser und Kohlenfaser) hergestellt werden. (Sub-EKD)
	**führen** – ein einfaches Muster für Kausalbeziehungen, z. B.: Die Rotation der Erde trägt ebenfalls zur Verwirbelung der Luftmassen bei, und die Schiefstellung der Rotationsachse der Erde zur Ebene, die die Erdbahn durch das Umkreisen der Sonne bildet, (ekliptikale Ebene) führt zu jahreszeitlichen Luftströmungen. (sic!) (Sub-EKD)

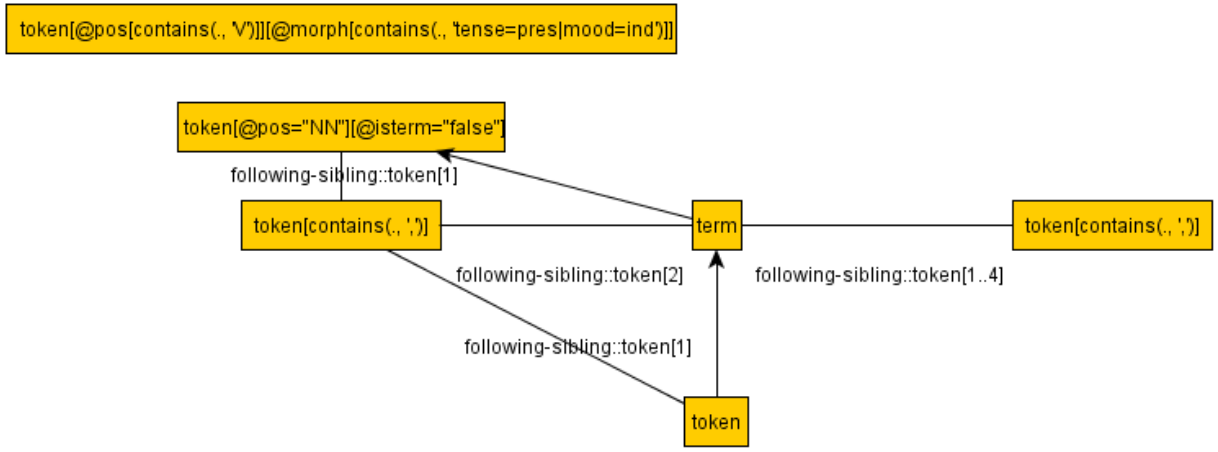
	**ausdrücken2** – ein wahrscheinlich eher seltenes Muster mit Modalverb, z. B.: Der aerodynamische Wirkungsgrad einer Anlage kann über das Verhältnis des Leistungsbeiwertes der Maschine zum Betzschen Leistungsbeiwert ausgedrückt werden und liegt demnach bei etwa 70 % bis 85 % je nach Windverhältnissen und Auslegung. (Sub-EKD)
	**nutzen4** – eine modale und passivierte Variante für das *knowledge pattern* „nutzen“, z. B.: Das Drosselventil kann auch zum vollständigen Abschalten des Rotors genutzt werden. (Sub-EKD)

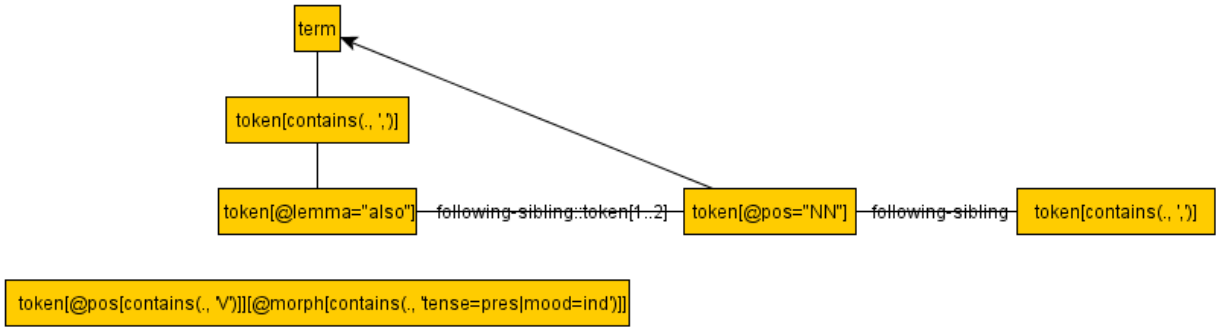


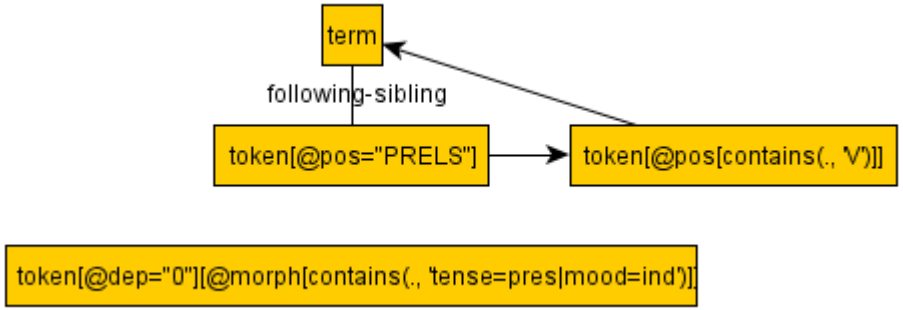
Apposition1 – dieses Muster extrahiert die Sequenz „Terminus + , + Artikel oder Adjektiv + Terminus (maximal ein *token* zwischen dem Artikel oder Adjektiv und dem zweiten Terminus) + ,“, zudem muss ein Verb im Präsens Indikativ im Satz vorhanden sein, z. B.:
In der Gondel, dem Maschinenhaus eines riesigen Windkraftrads, wird die Drehbewegung der Rotorblätter aufgenommen und über Induktionsspulen in der Turbine in fließenden Strom verwandelt. (Sub-EKD)



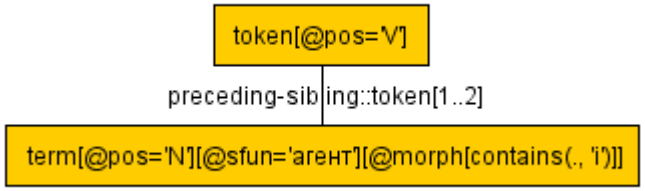
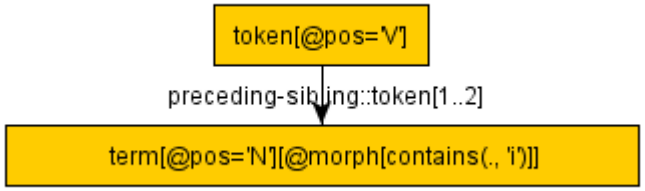
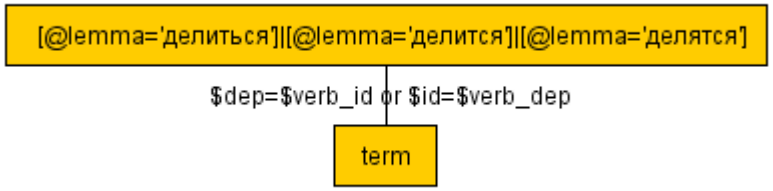
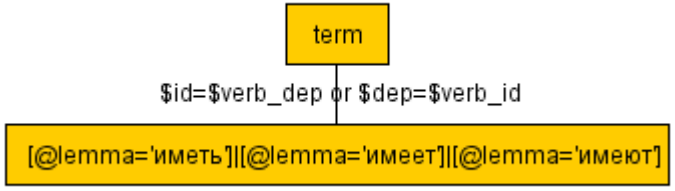
Apposition2 – ein möglicherweise redundantes Muster zur Identifikation von Appositionen, das aber auch andere KRCs extrahiert, z. B.:
Im zylindrischen Schacht 10 ist die vertikale Rotoranlage angeordnet, bestehend aus dem Rotor 5, der Rotorachse 4, welche gleichzeitig den Generator 6 aufnimmt und den Etagenböden 7. (Sub-EKD)

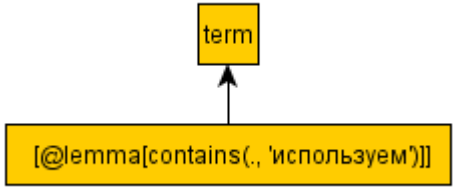
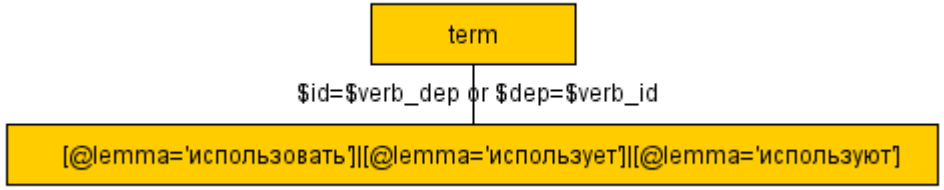
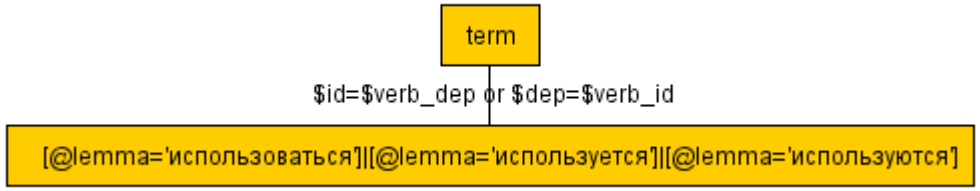
<table>
<tr>
<td> 
</td>
<td>Apposition3 – dieses Muster extrahiert die Sequenz „Nomen (kein Terminus) + , + token + Terminus + , (nach maximal 4 tokens)“, zudem muss ein Verb im Präsens Indikativ im Satz vorhanden sein, z. B.:
Bei einer Betrachtung der vertikalen Unterteilung der Atmosphäre ist alleine deren untere Schicht, die <u>Troposphäre</u>, für eine Nutzung der Windenergie von Interesse. (Sub-EKD)

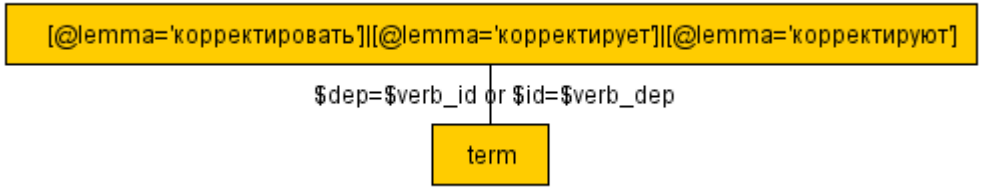
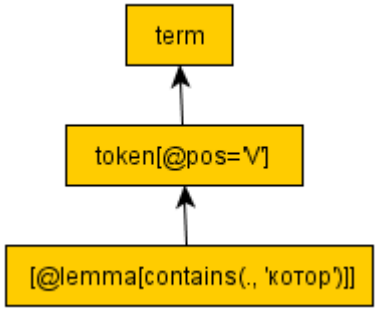
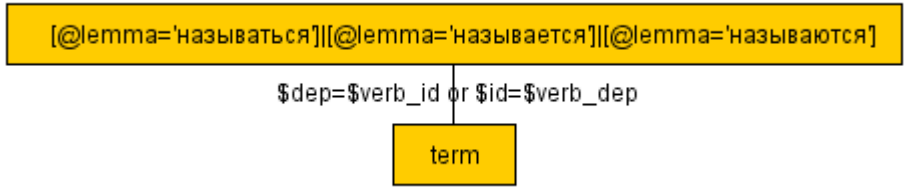
<u>Gasgefüllte Röhren</u> sind zum Beispiel Quecksilberdampfgleichrichter , gasgefüllte Photozellen , Nixieröhren , Glimmstabilisatoren und Thyratrons . (sic!) (Sub-TKD)</td>
</tr>
<tr>
<td> 
</td>
<td>also – auch dieses Muster extrahiert eine token-Sequenz, z. B.:
<u>Infraschall</u>, also Schall mit Frequenzen zwischen 0 Hz und 16 Hz, tritt z.B. im Zusammenhang mit Erdbeben oder tektonischen Verschiebungen als seismische Welle auf. (Sub-EKD)</td>
</tr>
</table>

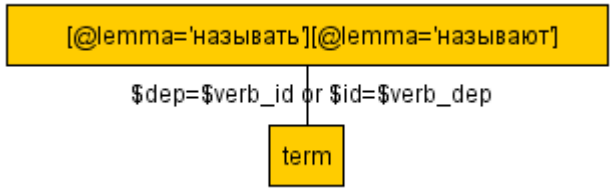
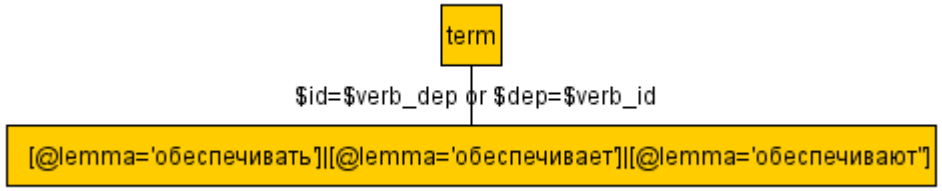
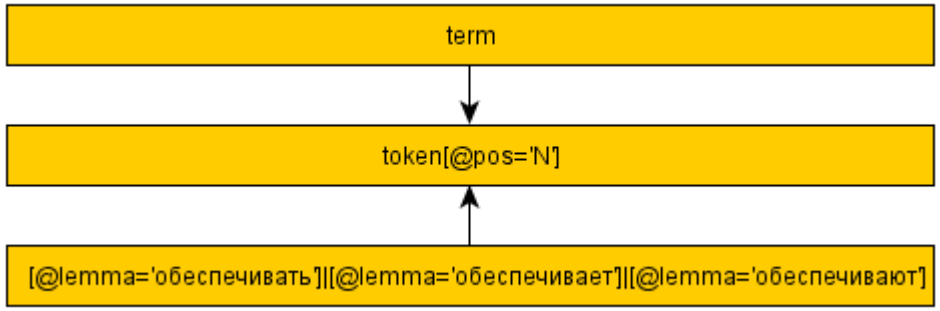
	**Relativ** – dieses Muster sucht auf relativ einfache Art und Weise nach Relativsätzen, die sich auf einen Terminus beziehen und in einen Hauptsatz eingebettet sind, z. B.: Bei großen Maschinen wird der Erregerstrom von einem Gleichstromgenerator, Erregermaschine genannt, geliefert, die unmittelbar oder über ein Getriebe mit der Welle der Synchronmaschine gekuppelt ist. (Sub-TKD)

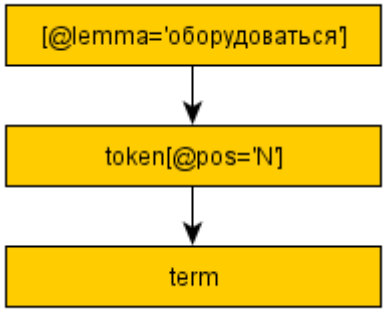
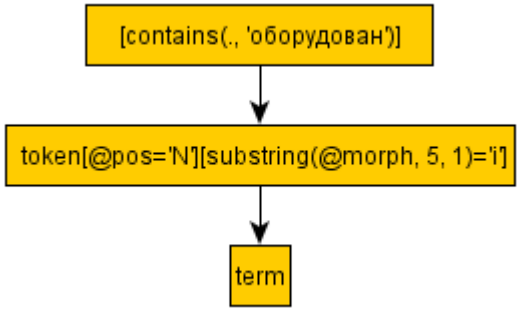
Anhang 11: Genutzte Extraktionsmuster und durch sie extrahierte KRCs – Russisch

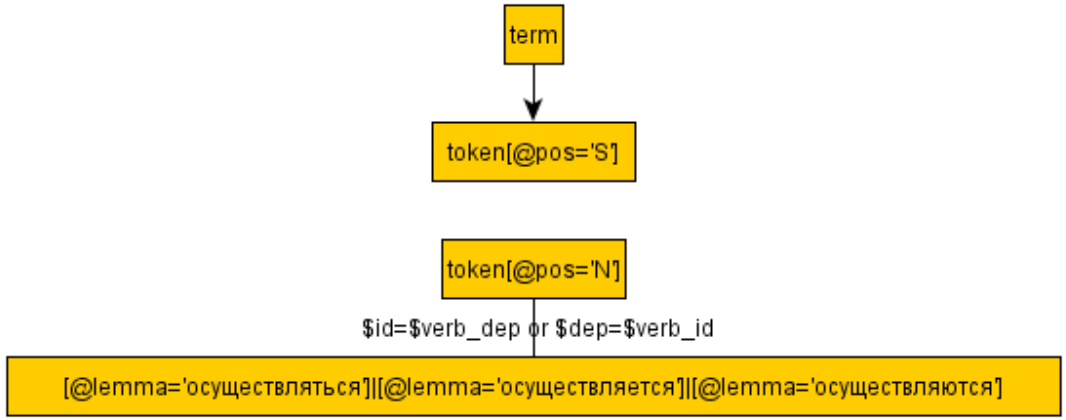
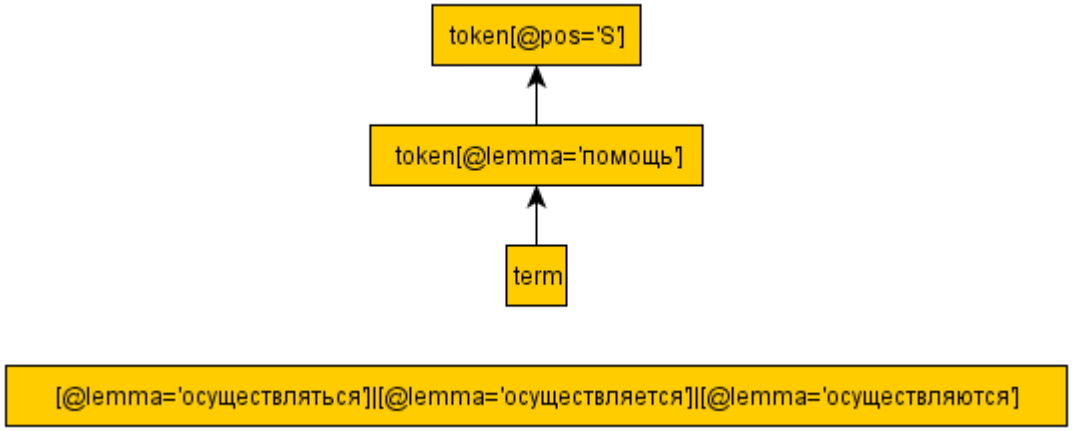
	**agent1** – dieses wahrscheinlich nicht sehr präzise Muster sucht nach Termini im Instrumental, die als Agens annotiert wurden und denen ein Verb unmittelbar voraus geht, z. B.: Топливо подается <u>топливным насосом</u> под давлением в топливный коллектор. (Sub-EKR)
	**agent2** – eine nur unwesentlich modifizierte Variante desselben Musters, z. B.: Существует два способа зажигания <u>дуги</u> покрытыми электродами: впритык и скольжением, чирканьем. (Sub-TKR)
	**делиться** – ein sehr einfaches Muster, das auf dem *knowledge pattern* „делиться“ („sich teilen“, „sich aufteilen“) basiert, z. B.: <u>Рабочий контур</u>, согласно требованиям ЕЭС, должен делиться на основной и вспомогательный. (Sub-EKR)
	**иметь1** – ein einfaches Muster für meronymische Beziehungen mit dem *knowledge pattern* „иметь“ („haben“), z. B.: <u>Стояночная тормозная система</u> имеет механический привод, как правило, на задние колеса. (Sub-EKR)

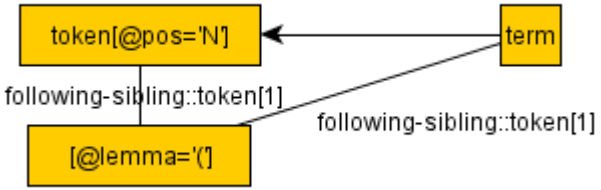
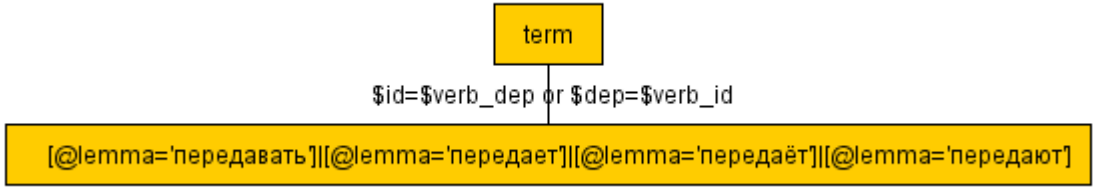
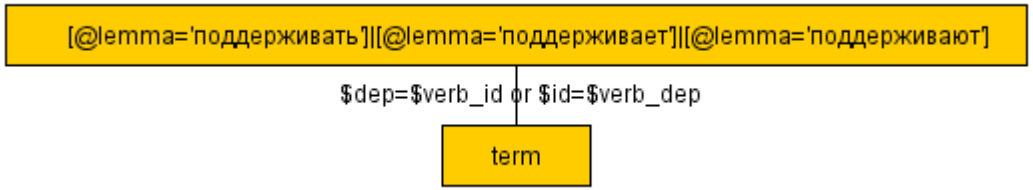
	**использовать1** - ein Muster auf Grundlage des *knowledge patterns* „использовать“ („benutzen“) in der Form als Partizip Präsens Passiv, z. B.: <u>Водяные радиаторы</u>, используемые для охлаждения наддувочного воздуха и устанавливаемые в системах с автономным жидкостным низкотемпературным контуром (рис. 11, а), работают так же, как основные радиаторы охлаждения воды. (Sub-EKR)
	**использовать 2** – dasselbe Muster als regulär konjugiertes Verb, z. B.: <u>Систему жидкостного охлаждения</u> наиболее целесообразно использовать в форсированных двигателях и в двигателях с относительно большим рабочим объемом цилиндра; <u>систему воздушного охлаждения</u> – в двигателях с рабочим объемом цилиндра до 1 л независимо от степени форсировки и в двигателях с небольшой литровой мощностью. (Sub-EKR)
	**использоваться** – eine Abwandlung desselben Musters als lexikalisierte Passivform, z. B.: На легковых автомобилях большого класса часто используются дисковые <u>тормозные механизмы</u> на передних колесах и барабанные колодочные на задних колесах. (Sub-EKR)

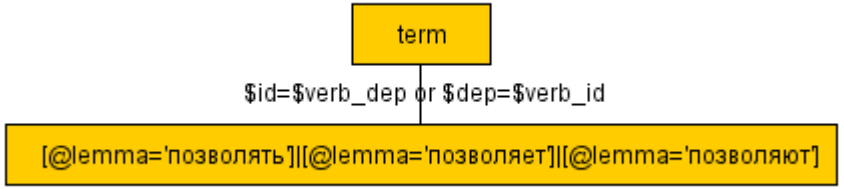
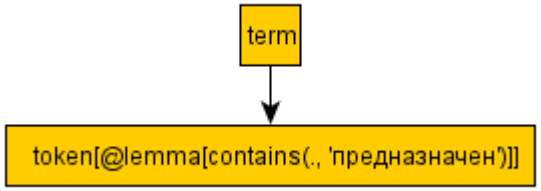
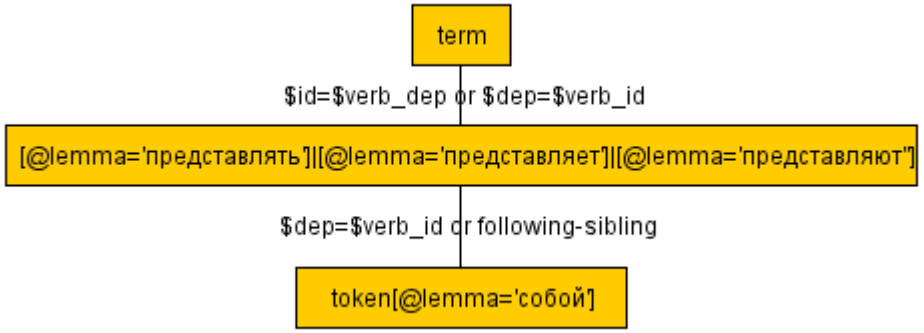
	**корректировать** – ein möglicherweise fachgebietsspezifisches Muster mit dem *knowledge pattern* „корректировать“ („korrigieren“), z. B.: На основе сигнала датчика электронный блок управления двигателем корректирует длительность сигнала открытия форсунки и угол опережения зажигания. (Sub-EKR)
	**который** – ein Muster zur Abfrage von Relativsätzen, das mitunter auch überraschende Ergebnisse erbringt, z. B.: Спокойная сталь - это сталь, выдержанная некоторое время в ковшах вместе с раскислителями (кремний, марганец, алюминий), которые соединяясь с растворенным кислородом, превращаются в оксиды и выплывают на поверхность массы стали. (Sub-TKR)
	**называться** – ein einfaches Muster auf Grundlage des *knowledge patterns* „называться“ („genannt werden“), z. B.: Верхней мертвой точкой (ВМТ) называется наивысшая точка хода поршня в своем цилиндре. (Sub-EKR)

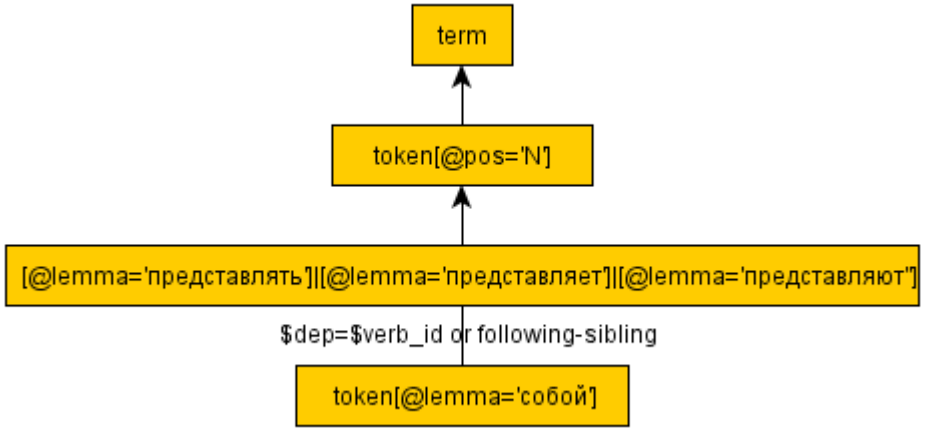
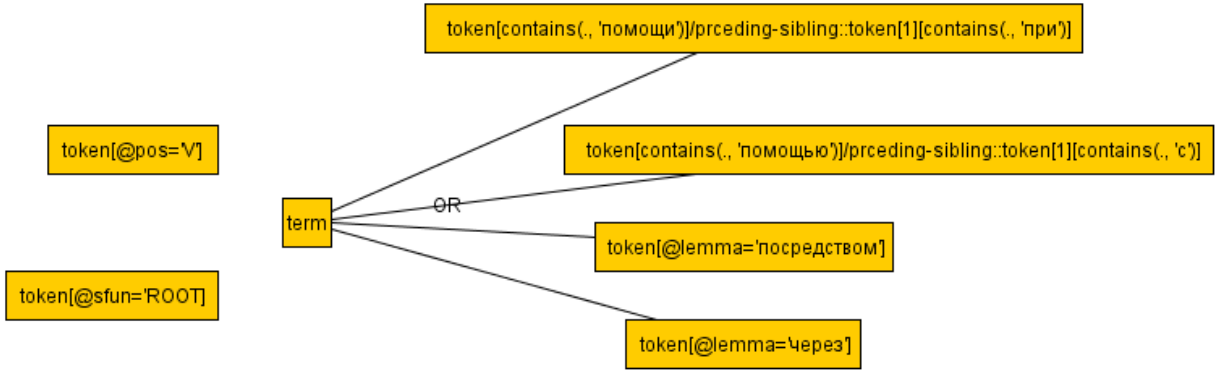
	**называть** – die Grundform desselben Musters ohne Passivierung, die aber insbesondere auch unbestimmt-persönliche Sätze einschließt, z. B.: Стыковую сварку с разогревом стыка до пластического состояния и последующей осадкой называют - сваркой оплавлением. (Sub-TKR)
	**обеспечивать1 -** ein Muster auf Grundlage des *knowledge patterns* „обеспечивать“ („gewährleisten“, „sicherstellen“), z. B.: Рулевой привод обеспечивает правильное соотношение углов поворота управляемых колёс. (Sub-EKR)
	**обеспечивать 3** – eine Variante desselben Musters mit abhängigem Definiendum, z. B.: Один контур гидропривода обеспечивает работу правого переднего и левого заднего тормозных механизмов, другой — левого переднего и правого заднего. (Sub-EKR)

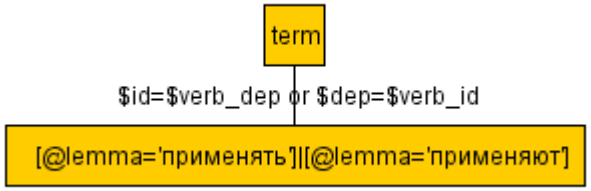
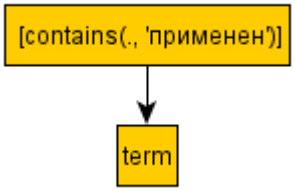
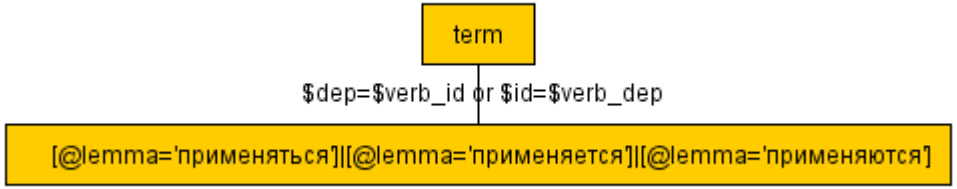
	**оборудоваться** – ein Muster auf Grundlage der lexikalisierten Passivform „оборудоваться“ („ausgerüstet werden“, „ausgestattet werden“), z. B.: Современный автомобиль оборудуется рабочей, запасной, стояночной и вспомогательной тормозными системами. (Sub-EKR)
	**оборудован** – dieses Muster sucht Varianten des *knowledge patterns* „оборудовать“ („ausrüsten“), das hier als Partizip Präsens Passiv fungiert, z. B.: Кроме того, антиблокировочная система оборудована системой самодиагностики, выявляющей неисправности компонентов. (Sub-EKR)

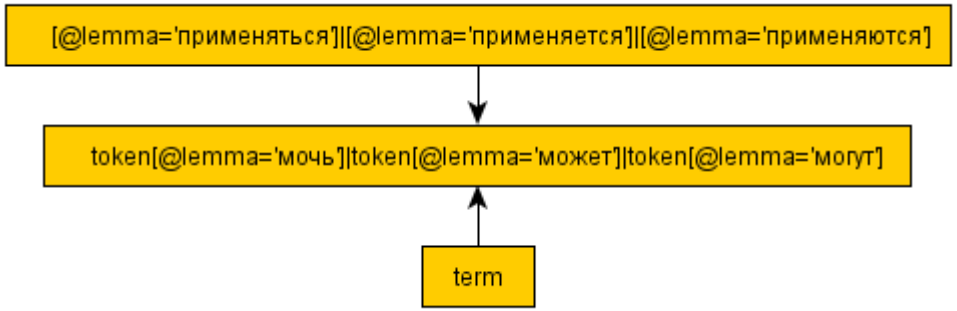
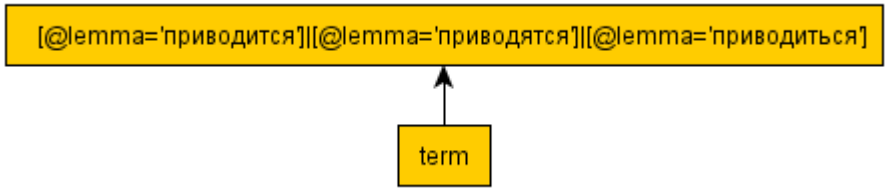
	**осуществляться1** – ein Muster auf Grundlage des *knowledge patterns* „осуществляться“ („erfolgen“, „ausgeführt werden“), in dem das Definiendum als Präpositionalphrase realisiert ist, z. B.: В <u>термосифонной системе охлаждения</u> циркуляция осуществляется за счет разницы в плотности холодной и горячей жидкости. (Sub-EKR)
	**осуществляться2** – eine Variante desselben Musters mit konkretisierter Präpositionalphrase, z. B.: Вывод из блока двигателя картерных газов осуществляется при помощи <u>системы вентиляции</u> полузакрытого типа. (Sub-EKR)

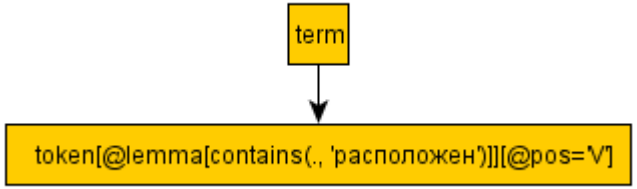
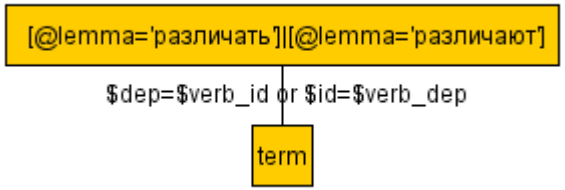
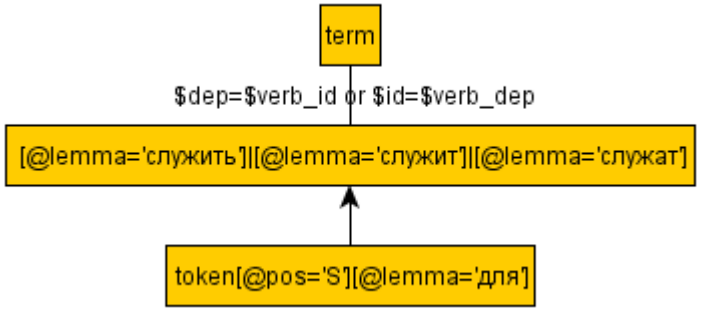
	**parenthese1** – ein einfaches sequentielles Muster, in dem ein Nomen von einer Parenthese gefolgt wird, die einen Terminus enthält, z. B.: Антиблокировочная система (АБС) — это система, предотвращающая блокирование колес при торможении, т.к. при блокировке колес автомобиль полностью теряет управляемость. (Sub-EKR)
	**передавать** – ein möglicherweise fachgebietsspezifisches Muster mit dem *knowledge pattern* „передавать“ („übertragen“), z. B.: Дифференциал передаёт крутящий момент от главной передачи к полуосям и позволяет им вращаться с разной скоростью при повороте автомобиля и на неровностях дороги. (Sub-EKR)
	**поддерживать** – ein möglicherweise fachgebietsspezifisches Muster auf Grundlage des *knowledge patterns* „поддерживать“ („unterstützen“), z. B.: Необходимую температуру жидкости в системе охлаждения автоматически поддерживает термостат. (Sub-EKR)

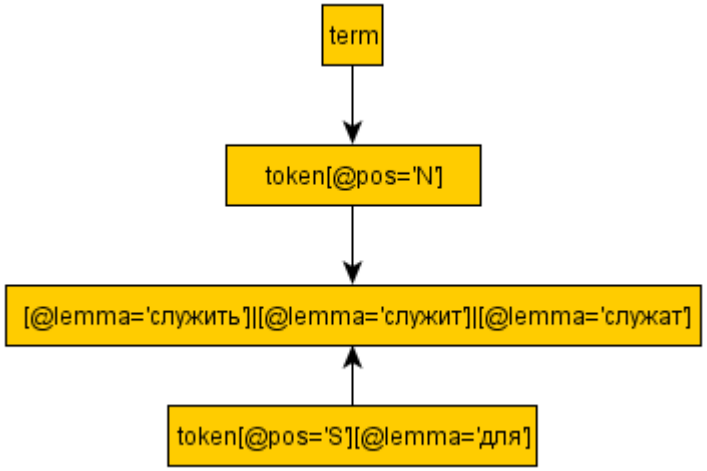
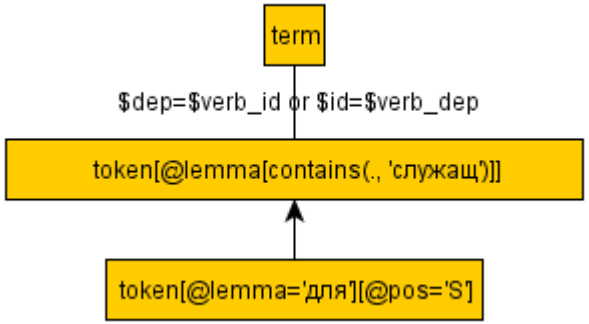
	**позволять** – ein ebenfalls einfaches Muster auf Grundlage des *knowledge patterns* „позволять“ („erlauben“, „ermöglichen“), z. B.: <u>Карданная передача</u> позволяет передавать крутящий момент между валами, расположенными под изменяющимся при движении автобуса углом. (Sub-EKR)
	**предназначен** – ein Muster mit dem *knowledge pattern* „предназначен“ („vorgesehen für“), z. B.: <u>Сцепление</u> предназначено для кратковременного разъединения двигателя от трансмиссии и плавного их соединения в тех случаях, когда требуется затормозить автомобиль, начать движение или переключить передачи. (Sub-EKR)
	**представлять1** – ein einfaches Muster auf Grundlage des *knowledge patterns* „представлять собой“ („darstellen“), z. B.: <u>Зубодолбёжный долбяк</u> представляет собой зубчатое колесо, снабжённое режущими кромками. (Sub-TKR)

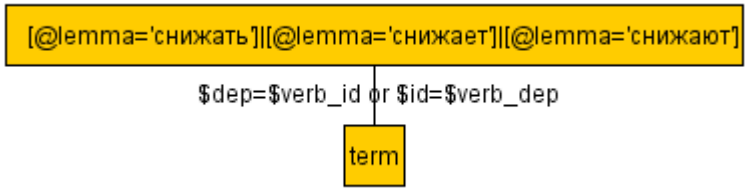
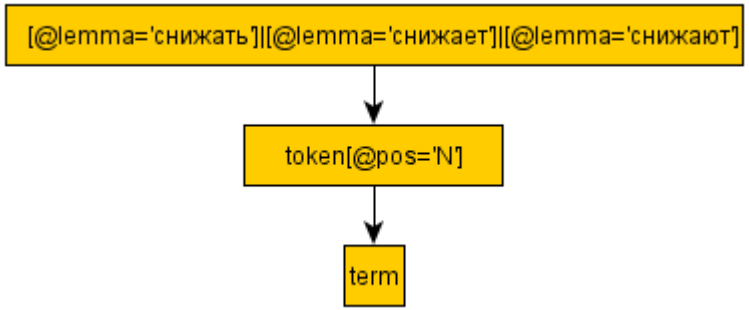
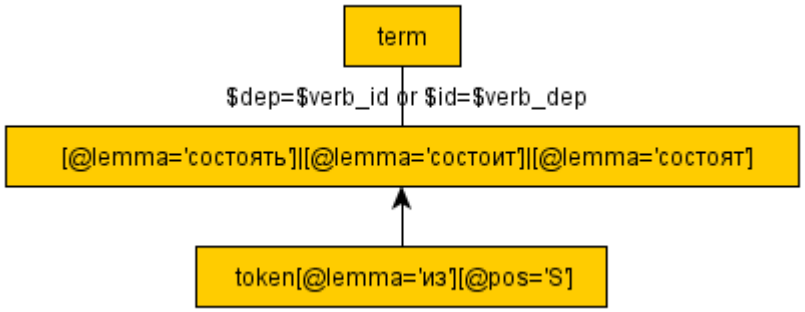
	**представлять 2** – eine leichte Variation desselben Musters zum Abfangen variierender Parser-Ausgaben, z. B.: Лауэграмма представляет собой систему пятен – дифракционных максимумов (которые еще называют рефлексами), по расположению которых можно судить о внутренней структуре и ориентировке кристалла. (Sub-TKR)
	**prep1** – dieses wahrscheinlich eher unsaubere Muster knüpft die Nennung eines Terms an bestimmte Präpositionen, die in einen vollständigen Hauptsatz eingebettet sein müssen, z. B.: Рулевой механизм предназначен для передачи изменения угла поворота рулевого колеса через рулевой привод на управляемые колёса. (Sub-EKR)

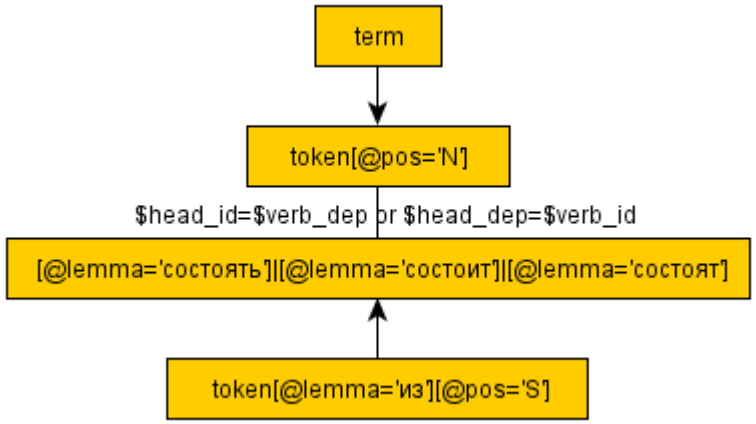
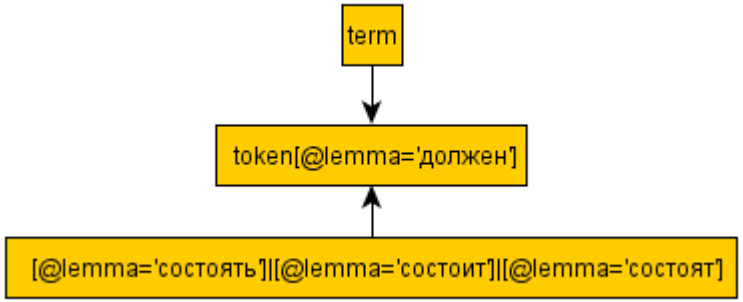
	**применять1** – ein einfaches Muster auf Grundlage des *knowledge patterns* „применять“ („anwenden“, „einsetzen“), z. B.: При дуговой сварке применяют неплавящиеся и плавящиеся электроды. (Sub-TKR)
	**применять 2** – eine Variante desselben Musters, in der das Verb ähnlich einem deutschen Zustandspassiv eingesetzt wird, z. B.: На автомобильных двигателях применены термостаты с жидкостным и твердым наполнителями. (Sub-EKR)
	**применяться1** – ein einfaches Muster mit der lexikalisierten Passivform „применяться“ („angewandt werden“, „verwendet werden“), z. B.: Щитки и маски (рис.2) применяются для предохранения глаз и кожи лица сварщика от вредного влияния ультрафиолетовых лучей и брызг расплавленного металла. (Sub-TKR)

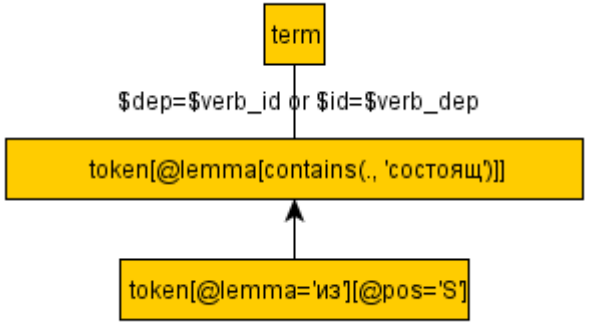
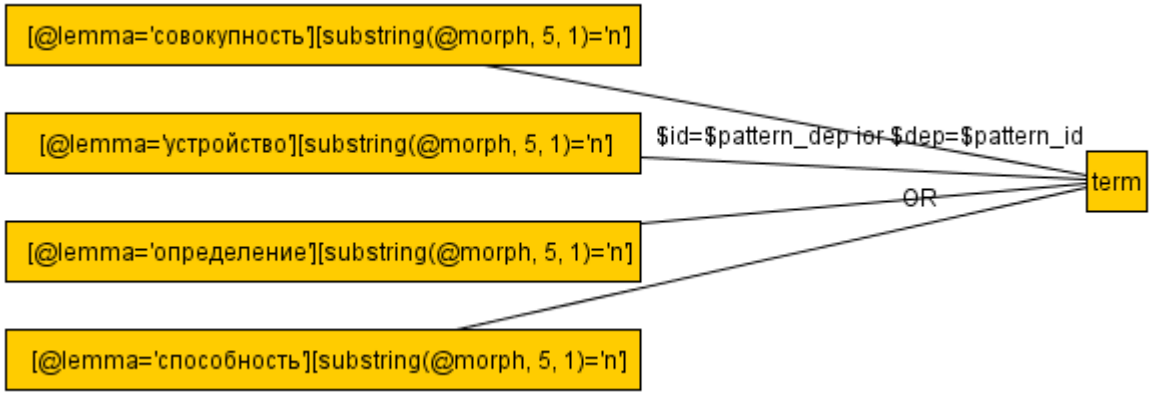
	**применяться2** – eine Variante desselben Musters mit Modalverb, z. B.: <u>Барабанные тормоза</u> могут применяться как на передних, так и на задних колесах, хотя из-за относительно низкой эффективности с 1970-х их устанавливают в большинстве случаев только на задней оси. (Sub-EKR)
	**приводиться** – ein einfaches Muster auf Grundlage des *knowledge patterns* „приводиться“ („angetrieben werden“), das wahrscheinlich fachgebietsspezifisch ist, z. B.: <u>Рабочая тормозная система</u> приводится в действие нажатием на педаль тормоза, которая располагается в ногах у водителя (исключение — автомобили для обучения принципам вождения, дополнительная группа педалей располагается в ногах у инструктора, а также нередко — модели, предназначенные для использования инвалидами, или переоборудованные для них). (Sub-EKR)

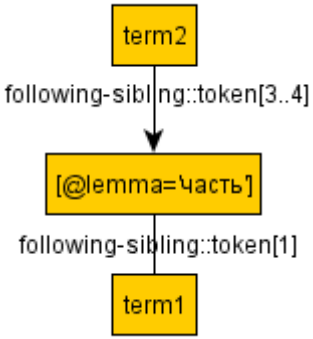
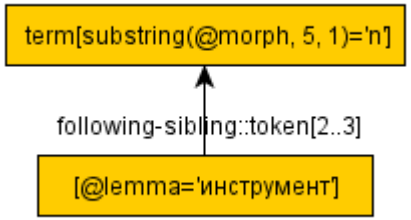
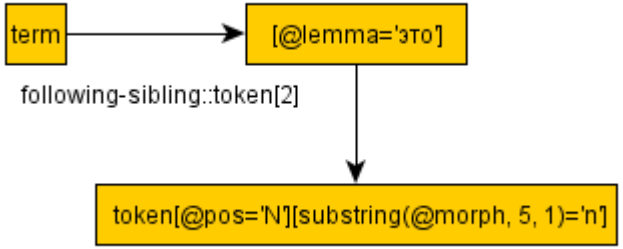
	**расположен** – ein einfaches Muster zur Extraktion von Lagebeziehungen mit dem *knowledge pattern* „расположен“ („gelegen sein“, „sich befinden“), z. B.: Трансмиссионная сборка расположена слева от двигателя и передает развиваемый последним крутящий момент на передние колеса автомобиля. (Sub-EKR)
	**различать** – ein einfaches Muster mit dem *knowledge pattern* „различать“ („unterscheiden“), das auch unbestimmt-persönliche Sätze abdeckt, z. B.: По способу заполнения швов по сечению различают однопроходные, однослойные швы, многопроходные и многослойные. (Sub-TKR)
	**служить1** – ein einfaches Muster mit dem *knowledge pattern* „служить“ („dienen“) für die Extraktion funktionaler Beziehungen, z. B.: Регулятор давления служит для регулирования давления в гидроприводе тормозных механизмов задних колёс в зависимости от нагрузки на заднюю ось автомобиля. (Sub-EKR)

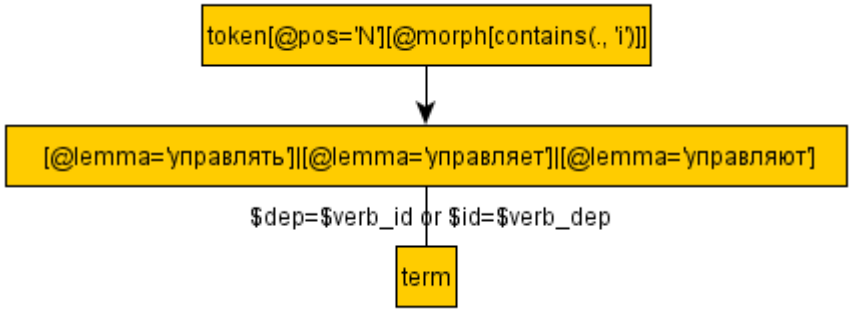
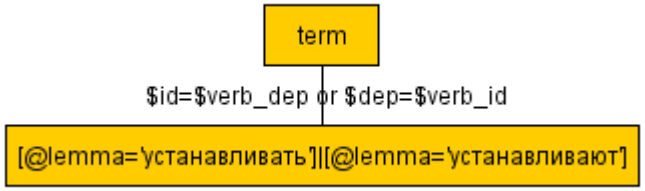
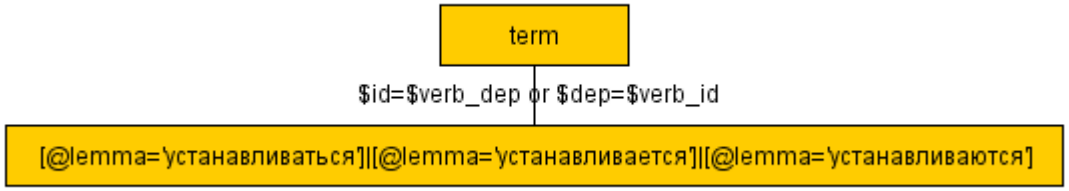
	**служить3** – eine wahrscheinlich redundante Variante desselben Musters mit abhängigem Definiendum, das allerdings auch komplexe Definienda identifiziert, z. B.: Тормозная система служит для снижения скорости и быстрой остановки автомобиля, а также для удержания его на месте при стоянке. (Sub-EKR)
	**служить5** – eine Variante desselben Muster, in der das Verb als Partizip Aktiv Präsens realisiert ist, z. B.: На автомобилях обязательно должны быть установлены: рабочая тормозная система, используемая при движении автомобиля для снижения скорости и полной остановки; стояночная тормозная система, служащая для удержания остановленного автомобиля на месте; запасная тормозная система, предназначена для остановки автомобиля при выходе из строя рабочей тормозной системы. (Sub-EKR)

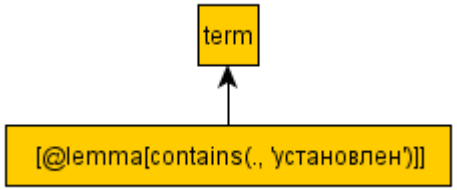
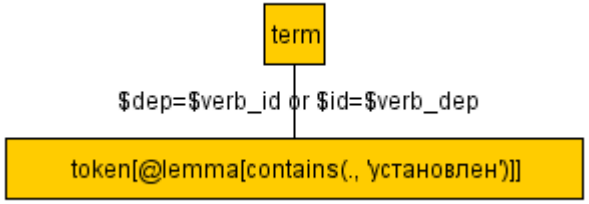
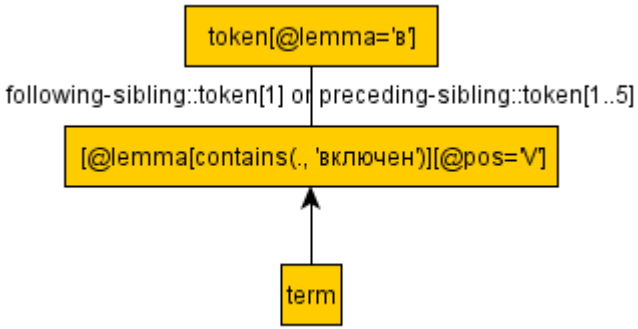
	**снижать1** – ein wahrscheinlich fachgebietsspezifisches Muster mit dem *knowledge pattern* „снижать“ („senken“, „reduzieren“), z. B.: <u>Вакуумный усилитель</u> тормозов снижает усилие на тормозной педали, облегчая тем самым управление автомобилем. (Sub-EKR)
	**снижать2** – eine wahrscheinlich redundante Variante desselben Musters zum Abfangen variierender Parser-Ausgaben
	**состоять1** – die Grundform des *knowledge patterns* „состоять“ („bestehen (aus)“) für die Extraktion meronymischer Beziehungen, z. B.: Каждая <u>тормозная система</u> состоит из тормозных механизмов, которые обеспечивают затормаживание колес или вал трансмиссий, и тормозного привода приводящего в действие тормозной механизм. (sic!) (Sub-EKR)

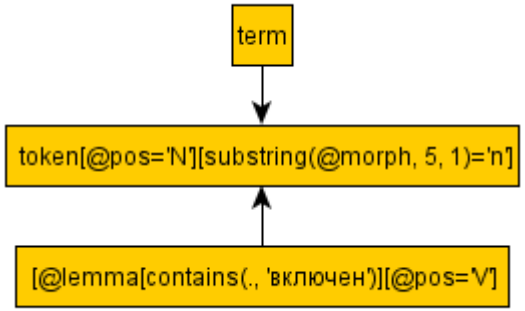
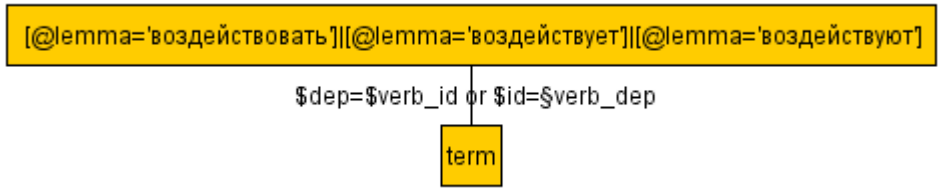
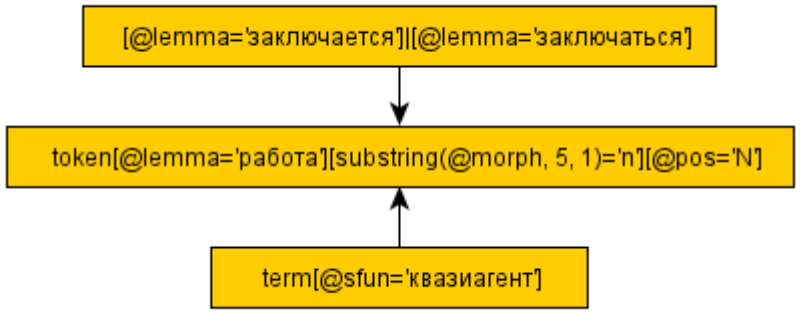
	**состоять3** – eine Variante desselben Musters mit abhängigem Definiendum, z. B.: <u>Пробка радиатора</u> состоит из корпуса, парового и воздушного клапанов и запорной пружины. (Sub-EKR)
	**состоять4** – eine Variation des Musters mit Modalverb, z. B.: Согласно этому стандарту <u>тормозное управление</u> должно состоять из четырех систем: рабочей, запасной, стояночной и вспомогательной. (Sub-EKR)

	**состоять5** – dasselbe Muster mit dem *knowledge pattern* als Partizip Präsens Aktiv, z. B.: Механический привод, состоящий из тяг и рычагов, применяют в основном в тормозных системах с ручным управлением (вспомогательная тормозная система -,,стояночный-тормоз''). (Sub-EKR)
	**tire1** – dieses sequentielle Muster sucht nach Kombinationen von Termini mit bestimmten nachfolgenden Schlüsselwörtern im Nominativ, die für die prädikatlosen Bindestrich-Konstruktionen typisch sind, z. B.: Виккерса метод — определение твердости материала вдавливанием в поверхность образца или изделия алмазного индентора (наконечника), имеющего форму правильной 4-гранной пирамиды с 2-гранным углом 136° при вершине. (Sub-EKR)

	**tire2** – dieses sequentielle Muster sucht nach zwei Termini, die durch eine meronymische Beziehung miteinander verbunden sind, z. B.: Основная часть волоки — волочильный глазок (из стали, твердых сплавов, алмазов), представляюший собой отверстие постепенно уменьшающегося сечения, через которое протягивается обрабатываемый металл. (Sub-EKR)
	**tire3** – auch dieses Muster ist recht spezifisch, indem es nur nach Abfolgen von Termini und einem bestimmten Schlüsselwort sucht, z. B.: Вороток — ручной инструмент для вращения режущих инструментов: метчиков, разверток, круглых плашек и т.п. (Sub-EKR)
	**tire4** – bei diesem Muster handelt es sich um die zweite Grundform der russischen KRCs mit Bindestrichkopula, z. B.: Колесные тормоза — это основные рабочие тормоза, обеспечивающие замедление движения и остановку машины. (Sub-EKR)

	**управлять** – ein Muster mit dem *knowledge pattern* „управлять“ („steuern“) sowie einem nominalen Objekt im Instrumental, z. B.: Антиблокировочная система тормозов (ABS) управляет гидравлическим давлением в приводе тормозов всех четырех колес при резком торможении и торможении на опасном дорожном покрытии, что предотвращает блокировку колес. (Sub-EKR)
	**устанавливать1** – ein einfaches Muster mit dem *knowledge pattern* „устанавливать“ („aufstellen“, „installieren“), das auch unbestimmt-persönliche Sätze einschließt, z. B.: На автомобили, не оснащенные ABS, для предотвращения заноса при интенсивном торможении устанавливают регуляторы давления в гидроприводах тормозных механизмов задних колес, смонтированные на главном тормозном цилиндре. (Sub-EKR)
	**устанавливаться** – ein einfaches Muster mit dem *knowledge pattern* „устанавливаться“ („aufgestellt werden“, „installiert werden“), z. B.: Двигатель устанавливается поперечно в переднюю часть автомобиля, слева к нему подсоединена трансмиссионная сборка, посредством которой крутящий момент передается на ведущие передние колеса. (Sub-EKR)

	**установить1** - ein Muster auf Grundlage des *knowledge patterns* „установить“ („aufstellen“, „installieren“), das hier als Partizip Präteritum Passiv eingesetzt wird, z. B.: Байпасный канал расположен в корпусе дроссельной заслонки параллельно каналу дроссельной заслонки в месте, где установлен сервопривод регулятора оборотов холостого хода. (Sub-EKR)
	**установить2** – eine leichte Modifikation desselben Musters, z. B.: Вакуумный усилитель (рисунок 9.3), установленный между механизмом педали и главным тормозным цилиндром, при торможении за счет разрежения во впускной трубе двигателя через шток и поршень первой камеры главного цилиндра создает дополнительное усилие, пропорциональное усилию от педали. (Sub-EKR)
	**включить1** – ein Muster auf Grundlage des *knowledge patterns* „включить“ („einschließen“, „beinhalten“), das hier als Partizip Präteritum Passiv fungiert, z. B.: В гидравлический привод тормозов включены вакуумный усилитель 6 и двухконтурный регулятор давления задних тормозов 9. (Sub-EKR)

	**включить2** – eine Variante desselben Musters, z. B.: <u>Регулятор давления</u> тормозов включен в оба контура тормозной системы, и через регулятор давления тормозов тормозная жидкость поступает к обоим задним тормозным механизмам. (Sub-EKR)
	**воздействовать** – ein einfaches – und möglicherweise fachgebietsspezifisches – Muster auf Grundlage des *knowledge patterns* „воздействовать“ („einwirken“, „betätigen“), z. B.: <u>Стояночный тормоз</u> воздействует через тросовые тяги на тормозные механизмы задних колес. (Sub-EKR)
	**заключаться** – dieses Muster sucht ebenfalls nach einem bestimmten Schlüsselwort, z. B.: Работа <u>системы тормозов</u> заключается в передаче усилия, прикладываемого к педали тормоза водителем, к тормозам колес. (Sub-EKR)

Tabellenverzeichnis

Tabelle Nr.	Titel	Seite	Abschnitt
1	Strukturtypen wissenshaltiger Kontexte und weiterer Kontexttypen nach Pearson (1998)	20-21	2.3.2.3 Linguistische Klassifikation *(wissenshaltiger Kontexte)*
2	Klassifikation wissenshaltiger Kontexte	22	2.3.2.3 Linguistische Klassifikation *(wissenshaltiger Kontexte)*
3	Existierende Typologien semantischer Relationen und eigene Zusammenfassung	26-27	2.3.3.2. Entwicklung einer Typologie semantischer Relationen für die computergestützte Terminologiearbeit und die Identifikation wissenshaltiger Kontexte
4	Aus der Literatur entnommene Elemente für knowledge patterns	38-42	3.2.3 Knowledge patterns
5	Überblick über ausgewählte Kriterien zur Beschreibung von Textkorpora nach Atkins et al. (1992)	73-74	4.1.2 Lexikographische Korpora und Web-Korpora
6	Überblick über die erstellten russischen und deutschen Textkorpora	76-77	4.1.3 Dokumentation der erstellten Korpora
7	Beispiel für eine Vierfeldertafel	80	4.1.4.1 Vergleich der Entwicklungs- und Testkorpora mit Referenzkorpora
8	Vergleichende statistische Charakterisierung der gesammelten Korpora	83	4.1.4.1 Vergleich der Entwicklungs- und Testkorpora mit Referenzkorpora
9	Wort- und Satzzahl in den Korpora sowie durchschnittliche Satzlänge	84	4.1.4.1 Vergleich der Entwicklungs- und Testkorpora mit Referenzkorpora
10	Top-10 terminologisch relevante Schlüsselwörter der russischen Korpora nach Analyse mit AntConc	85	4.1.4.2 Vergleich des Entwicklungskorpus‘ mit dem Testkorpus
11	Top-10 terminologisch relevante Schlüsselwörter der deutschen Korpora nach Analyse mit AntConc	85-86	4.1.4.2 Vergleich des Entwicklungskorpus‘ mit dem Testkorpus

12	Überblick über die von Annotator A ausgewählten KRC-Kandidaten für den ersten Annotationsdurchgang	87	4.2.1 Ausgangsdaten *(für die KRC-Annotation)*
13	Überblick über die von Annotator A ausgewählten KRC-Kandidaten für den zweiten Annotationsdurchgang	87	4.2.1 Ausgangsdaten *(für die KRC-Annotation)*
14	Beispiel für die Präsentation der KRC-Kandidaten während der Annotation des Goldstandards	88	4.2.2 Annotationsszenario und Annotation
15	Ergebnisse der deutschen Erstannotation pro Annotator	90	4.2.3.1 Deutsch *(Annotationsergebnisse)*
16	Vierfeldertafel für die Erstannotation der deutschen Korpora	90	4.2.3.1 Deutsch *(Annotationsergebnisse)*
17	Ergebnisse der deutschen Zweitannotation pro Annotator	91	4.2.3.1 Deutsch *(Annotationsergebnisse)*
18	Vierfeldertafel für die Zweitannotation der deutschen Korpora	91	4.2.3.1 Deutsch *(Annotationsergebnisse)*
19	Ergebnisse der russischen Erstannotation pro Annotator	91	4.2.3.2 Russisch *(Annotationsergebnisse)*
20	Vierfeldertafel für die Erstannotation der russischen Korpora	92	4.2.3.2 Russisch *(Annotationsergebnisse)*
21	Ergebnisse der russischen Zweitannotation pro Annotator	92	4.2.3.2 Russisch *(Annotationsergebnisse)*
22	Vierfeldertafel für die Zweitannotation der russischen Korpora	92	4.2.3.2 Russisch *(Annotationsergebnisse)*
23	Evaluation der Annotation	94	4.2.3.3 Quantitative Bewertung der Annotation
24	Ergebnis der qualitativen Analyse negativer Bewertungen für deutsche KRC-Kandidaten der Erstannotation	95-96	4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards
25	Ergebnis der qualitativen Analyse negativer Bewertungen für deutsche KRC-Kandidaten der Zweitannotation	97	4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards
26	Ergebnis der qualitativen Analyse negativer Bewertungen für russische KRC-Kandidaten der Erstannotation	98	4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards
27	Ergebnis der qualitativen Analyse negativer Bewertungen für russische KRC-Kandidaten der Zweitannotation	98-99	4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards

28	Ergebnis der qualitativen Analyse positiver Bewertungen für die untersuchten Korpora	99	4.2.3.4.2 Identifikation der falschen Positive
29	Größe des Goldstandards nach Erst- und Zweitannotation sowie der qualitativen Konsolidierung	100	4.2.3.4.3 Goldstandard
30	Unbekannte KRC-Kandidaten anhand von einfachen Zufallsstichproben aus allen vier Subkorpora	101	4.2.4 Abdeckung der Annotation
31	Klassifikation der in den Web-Korpora vorgefunden Texte	103-104	4.3.1 Theoriegeleitete Klassifikation von Textsorten in Web-Korpora
32	Überblick über Subkorpora des Goldstandards	106-107	4.3.2 Eigenschaften der Subkorpora
33	Fachgebietsverteilung in den Subkorpora des Goldstandards	108	4.3.2 Eigenschaften der Subkorpora
34	Evaluation für die Satzsegmentierung der russischen und deutschen Sub-Korpora	111	4.4.1 Vorverarbeitung *(der Subkorpora)*
35	Beispielsatz aus Sub-EKD nach Analyse mit den mate-Werkzeugen	115-116	4.4.2.2.1 Deutsch *(syntaktische Analyse)*
36	Zentrale syntaktische Relationen in SynTagRus	116-117	4.4.2.2.2 Russisch *(syntaktische Analyse)*
37	Beispielsatz aus Sub-EKR nach Analyse mit Hilfe der Annotationspipeline	118	4.4.2.2.2 Russisch *(syntaktische Analyse)*
38	Evaluation der Term-Erkennung	120	4.4.3 Formatierung *(der Subkorpora)*
39	Merkmalskatalog für die Untersuchung wissenshaltiger Kontexte	130	5.2.1.3 Systemisch-funktionale Linguistik
40	Unterschiede in den Wortartenfrequenzen des russischen Goldstandards, des RIC und der Extraktionskorpora	133-134	5.2.2.2.1 Russisch *(Wortartenfrequenzen)*
41	Unterschiede in den Wortartenfrequenzen des deutschen Goldstandards, des NCL und der Extraktionskorpora	136-137	5.2.2.2.2 Deutsch *(Wortartenfrequenzen)*
42	Kandidaten für russische Boostwörter	140-142	5.2.2.3.1 Russisch *(Boost- und Stopwörter)*
43	Kandidaten für russische Stopwörter	145-149	5.2.2.3.1 Russisch *(Boost- und Stopwörter)*
44	Kandidaten für deutsche Boostwörter	150-151	5.2.2.3.2 Deutsch *(Boost- und Stopwörter)*

45	Kandidaten für deutsche Stopwörter	154-155	5.2.2.3.2 Deutsch *(Boost- und Stopwörter)*
46	Semantische Klassen einfacher Prädikate in den KRCs des Goldstandards sowie Nicht-KRCs für das Sub-EKD	172	c) Ergebnisse *(Semantische Typen von „knowledge patterns“)*
47	Quantitative Typologie einfacher Prädikate in KRCs und Nicht-KRCs aus dem Sub-EKD	173	c) Ergebnisse *(Semantische Typen von „knowledge patterns“)*
48	Semantische Klassen einfacher Prädikate in den KRCs des Goldstandards sowie Nicht-KRCs für das Sub-EKR	174	c) Ergebnisse *(Semantische Typen von „knowledge patterns“)*
49	Quantitative Typologie einfacher Prädikate in KRCs und Nicht-KRCs aus dem Sub-EKR	175	c) Ergebnisse *(Semantische Typen von „knowledge patterns“)*
50	Ergebnisse der kontrastiven Analyse morphologischer Informationen für russische Verben der Goldstandard-KRCs im Vergleich zu Referenzdaten	180	5.2.3.2.1 Russisch (*Verbformenanalyse)*
51	Beispiele für Realisierungsformen russischer knowledge patterns	182-183	5.2.3.2.1 Russisch (*Verbformenanalyse)*
52	Ergebnisse der kontrastiven Analyse morphologischer Informationen für deutsche Verben der Goldstandard-KRCs im Vergleich zu Referenzdaten	184	5.2.3.2.2 Deutsch (*Verbformenanalyse)*
53	Beispiele für Realisierungsformen deutscher knowledge patterns	185-186	5.2.3.2.2 Deutsch (*Verbformenanalyse)*
54	Realisierung der Satzanfänge in den 292 KRCs des Sub-EKR	188	5.2.3.3.1 Russisch *(Realisierung des Definiendums)*
55	Realisierung der Satzanfänge in den 337 KRCs des Sub-EKD	191-192	5.2.3.3.2 Deutsch *(Realisierung des Definiendums)*
56	Analyse der maximalen syntaktischen Einbettungstiefe für deutsche KRCs im Vergleich zu Referenzdaten	195	5.2.3.4 Syntaktische Komplexität
57	Analyse der Satzlänge für russische KRCs im Vergleich zu Referenzdaten	195	5.2.3.5.1 Russisch *(Satzlänge)*
58	Analyse der Satzlänge für deutsche KRCs im Vergleich zu Referenzdaten	196	5.2.3.5.2 Deutsch *(Satzlänge)*

59	Häufigkeitsvergleich automatisch extrahierter Präpositionalphrasen für den russischen Goldstandard und die Zufallsstichprobe aus den Extraktionskorpora	197-198	5.2.3.6.1 Russisch *(Phraseologie)*
60	Häufigkeitsvergleich automatisch extrahierter Präpositionalphrasen für den deutschen Goldstandard und die Zufallsstichprobe aus den Extraktionskorpora	199	5.2.3.6.2 Deutsch *(Phraseologie)*
61	Merkmale für die multivariate Analyse deutscher KRCs	203-205	5.3.3 Merkmale und Daten *(multivariate Analyse deutscher KRCs)*
62	Merkmalsvektor für einen Beispielsatz	205-206	5.3.3 Merkmale und Daten *(multivariate Analyse deutscher KRCs)*
63	Regressionskoeffizienten, Standardfehler und Konfidenzintervalle für die Modellparameter des AIC-Modells	209	5.3.4.3 Interpretation der Regressionskoeffizienten
64	Ergebnisse der statistischen Merkmalsbewertung	210-211	5.3.4.3 Interpretation der Regressionskoeffizienten
65	Performanz des Baseline-Extraktionssystems auf den deutschen Goldstandard-Korpora	223	6.2.3.1.1 Baseline *(musterbasierte Extraktionsergebnisse Deutsch)*
66	Präzision der deutschen Extraktionsmuster auf den Entwicklungs- und Testdaten	224-225	6.2.3.1.1 Baseline *(musterbasierte Extraktion Deutsch)*
67	Performanz des Extraktionssystems auf den deutschen Goldstandard-Korpora nach Präzisionsoptimierung	227	6.2.3.1.2 Punktuelle Präzisionsverbesserung *(musterbasierte Extraktion Deutsch)*
68	Performanz des deutschen Extraktionssystems mit zusätzlichen Suchmustern aus der Literatur	229	6.2.3.1.3 Möglichkeiten der punktuellen Abdeckungsverbesserung *(musterbasierte Extraktion Deutsch)*
69	Ergebnisse der Maßnahmen zur Verbesserung der Abdeckung für das deutsche Extraktionssystem	231	6.2.3.1.3 Möglichkeiten der punktuellen Abdeckungsverbesserung *(musterbasierte Extraktion Deutsch)*
70	Performanz des Baseline-Extraktionssystems auf den russischen Goldstandard-Korpora	232	6.2.3.2.1 Baseline *(musterbasierte Extraktion Russisch)*
71	Performanz des russischen Extraktionssystems auf dem Sub-TKR nach verbesserter Term-Erkennung	235	6.2.3.2.2 Term-Erkennung *(musterbasierte Extraktion Russisch)*

72	Performanz des russischen Extraktionssystems mit zusätzlichen theoriebasierten Extraktionsmustern	236	6.2.3.2.3 Zusätzliche Extraktionsmuster *(musterbasierte Extraktion Russisch)*
73	Performanz des russischen Extraktionssystems nach Präzisionssteigerung	237	6.2.3.2.4 Punktuelle Präzisionsverbesserung *(musterbasierte Extraktion Russisch)*
74	Präzision der individuellen Suchmuster auf den russischen Korpora	238-239	6.2.3.2.4 Punktuelle Präzisionsverbesserung *(musterbasierte Extraktion Russisch)*
75	Merkmale für die Klassifikation russischer KRC-Kandidaten	244-245	6.3.1.1.2 Daten und Modellierung *(logistische Vorsortierung der Korpussätze Russisch)*
76	Merkmalsvektor für einen russischen Beispielsatz	245-246	6.3.1.1.2 Daten und Modellierung *(logistische Vorsortierung der Korpussätze Russisch)*
77	Ergebnisse der musterbasierten Extraktion nach Klassifikation der Korpussätze des Sub-TKR mit logistischer Regression in verschiedenen settings	249	6.3.1.1.3 Ergebnisse *(logistische Vorsortierung der Korpussätze Russisch)*
78	Erweiterter Merkmalskatalog	250-251	6.3.1.2.2 Daten und Modellierung *(Einbezug der Musterinformation in die logistische Modellierung Russisch)*
79	Erweiterter Merkmalsvektor für einen Beispielsatz	252	6.3.1.2.2 Daten und Modellierung *(Einbezug der Musterinformation in die logistische Modellierung Russisch)*
80	Modellparameter für ein u. a. um musterbezogene Informationen erweitertes Regressionsmodell	253	6.3.1.2.2 Daten und Modellierung *(Einbezug der Musterinformation in die logistische Modellierung Russisch)*
81	Auswirkungen der systematischen Variation von Merkmalswerten auf einer Zufallsstichprobe aus dem Sub-TKR	254-255	6.3.1.2.3 Ergebnisse *(Einbezug der Musterinformation in die logistische Modellierung Russisch)*
82	Merkmalsannotation für einen deutschen Beispielsatz	258-259	6.3.2.1 Daten und Modellierung *(Einbezug der Musterinformation in die logistische Modellierung Deutsch)*
83	Parameter des deutschen AIC-Modells	259	6.3.2.1 Daten und Modellierung *(Einbezug der Musterinformation in die logistische Modellierung Deutsch)*
84	Auswirkungen der systematischen Variation von Merkmalswerten auf einer Stichprobe aus dem Sub-TKD	261-262	6.3.2.2 Ergebnisse *(Einbezug der Musterinformation in die logistische Modellierung Deutsch)*

Abbildungsverzeichnis

Abbildung Nr.	Titel	Seite	Abschnitt
1	Kontexttypen und wissenshaltige Kontexte	13	2.2 Kontexttypen in der Linguistik und erste Beschreibung sowie Abgrenzung wissenshaltiger Kontexte von anderen Kontexttypen
2	Semantisch-sprachliches Modell wissenshaltiger Kontexte	15	2.3.1 Inhaltliche Definition (wissenshaltiger Kontexte)
3	Schematische Darstellung der Typologie semantischer Relationen nach Wüster/Nuopponen	24	2.3.3.2 Entwicklung einer Typologie semantischer Relationen für die computergestützte Terminologiearbeit und die Identifikation wissenshaltiger Kontexte
4	Einflussgrößen bei der Annotation definitorischer Aussagen nach Cramer (2011: 182)	45	3.3.1 Durchführbarkeit der Annotation und Einflussfaktoren
5	Annotationsexperimente nach Cramer (2011: 136)	48	3.3.2.1 Annotationsverfahren
6	Word-class lattice für aristotelische Definitionen nach Navigli und Velardi (2010)	59	3.4.2.1 Klassifikationsverfahren
7	Gewichtung der Ranking-Merkmale bei linearer Regression in verschiedenen Settings nach Walter (2010: 232)	67	3.4.3 Hybride Ansätze
8	Arbeitsschritte zur Erstellung des Goldstandards (qualitative Bewertung)	95	4.2.3.4 Qualitative Bewertung der Annotation und Erstellung des Goldstandards
9	Vorverarbeitung, linguistische Annotation und Formatierung der Korpora	109	4.4 Vorverarbeitung, Annotation und Formatierung der Korpora
10	mate-Annotations-Pipeline nach Björkelund et al. 2010: 34	113	4.4.2.1 Wortartentagging und Lemmatisierung
11	Tiger-Beispielsatz (Zinsmeister et al. 2001: 9)	115	4.4.2.2.1 Deutsch *(syntaktische Analyse)*
12	DTD für die Sub-Korpora	119	4.4.3 Formatierung *(der Subkorpora)*

13	Token-Annotation im XML-Format	120	4.4.3 Formatierung *(der Subkorpora)*
14	Relative Wortartenfrequenzen im russischen KRC-Goldstandard, in einer Stichprobe aus dem RIC sowie aus den Extraktionskorpora	133	5.2.2.2.1 Russisch *(Wortartenfrequenzen)*
15	Relative Wortartenfrequenzen im deutschen KRC-Goldstandard, in einer Stichprobe aus dem NCL sowie aus den Extraktionskorpora	136	5.2.2.2.2 Deutsch *(Wortartenfrequenzen)*
16	Bausteine für russische knowledge patterns in den KRCs des Sub-EKR	159	5.2.2.5.1 Russisch (*Bausteine für knowledge patterns)*
17	Bausteine für deutsche knowledge patterns in den KRCs des Sub-EKD	163	5.2.2.5.2 Deutsch (*Bausteine für knowledge patterns)*
18	Screenshot aus dem GermaNet Explorer für das Verb „einsetzen“	171	b) Methode *(Semantische Typen von „knowledge patterns“)*
19	Ausgabe des Stanford-Parsers	178	5.2.3.1 Methode *(Strukturelle und grammatische Merkmale)*
20	Algorithmus für die Extraktion von Präpositionalphrasen	179	5.2.3.1 Methode *(Strukturelle und grammatische Merkmale)*
21	Grammatische Realisierung von knowledge patterns in den KRCs des Sub-EKR	182	5.2.3.2.1 Russisch (*Verbformenanalyse)*
22	Grammatische Realisierung von knowledge patterns in den KRCs des Sub-EKD	185	5.2.3.2.2 Deutsch (*Verbformenanalyse)*
23	Realisierung des Definiendums in den 292 KRCs des Sub-EKR	187	5.2.3.3.1 Russisch *(Realisierung des Definiendums)*
24	Realisierung des Definiendums in den 337 KRCs des Sub-EKD	190	5.2.3.3.2 Deutsch *(Realisierung des Definiendums)*
25	QQ-Plot und Histogramm der studentisierten Devianzresiduen für das AIC-Modell	208	5.3.4.2 Evaluation der Modellanpassung
26	Präzisions-Abdeckungs-Kurve für ein einfaches Ranking deutscher KRC-Kandidaten aus dem Sub-EKD und dem Sub-TKD	212	5.3.4.4 Einfaches Klassifikationsmodell
27	XSLT-Regel für die Extraktion von Beispiel 97)	218	6.2.1 Implementierung der Extraktionsmuster

28	Einfache russische Extraktionsregel	218	6.2.1 Implementierung der Extraktionsmuster
29	Präzisions-Abdeckungs-Graph für das deutsche Baseline-Extraktionssystem	223	6.2.3.1.1 Baseline *(musterbasierte Extraktion Deutsch)*
30	Präzisions-Abdeckungs-Graph für das russische Baseline-Extraktionssystem	232	6.2.3.2.1 Baseline *(musterbasierte Extraktion Russisch)*
31	Vergleich der Präzisions-Abdeckungs-Graphen des russischen Baseline-Systems (_base) mit den entsprechenden Graphen nach der Präzisions- und Abdeckungsoptimierung (_extended_mod)	237	6.2.3.2.4 Punktuelle Präzisionsverbesserung *(musterbasierte Extraktion Russisch)*
32	f-Maße für die musterbasierten Extraktionssysteme auf den einzelnen Korpora vor (Baseline) und nach der Optimierung (End)	240	6.2.4 Diskussion *(musterbasierte Extraktion Russisch und Deutsch)*
33	Präzisions-Abdeckungs-Kurve für das Ranking russischer KRC-Kandidaten des Sub-TKR	248	6.3.1.1.3 Ergebnisse *(logistische Vorsortierung der Korpussätze Russisch)*
34	Präzisions-Abdeckungs-Kurve für das Ranking russischer KRC-Kandidaten des Sub-TKR mit dem erweiterten Regressionsmodell im Vergleich zu vorher erreichten Werten	254	6.3.1.2.3 Ergebnisse *(Einbezug der Musterinformation in die logistische Modellierung Russisch)*
35	Präzisions-Abdeckungs-Kurve für die Extraktion deutscher KRC-Kandidaten des Sub-TKD mit dem erweiterten Regressionsmodell im Vergleich zu zwei Baselines	260	6.3.2.2 Ergebnisse *(Einbezug der Musterinformation in die logistische Modellierung Deutsch)*

Abstract

Diese Arbeit widmet sich terminologischen Begriffsbeschreibungen, d. h. semantischen Beschreibungen fachspezifischer Begriffe in unterschiedlichen Arten von Gebrauchstexten. Sie bezieht ihre Legitimation aus dem Umstand, dass solche Beschreibungen insofern für den Transfer begriffsbezogenen Wissens eine unabdingbare Voraussetzung darstellen, als sie wesentliche Begriffsmerkmale benennen und den fraglichen Begriff zu anderen relevanten Begriffen in Beziehung setzen. Dies begründet die Wichtigkeit terminologischer Begriffsbeschreibungen – die hier als „wissenshaltige Kontexte“ bezeichnet werden – für wissensbasierte Tätigkeiten wie etwa die Translation, deren Blickwinkel im Wesentlichen eingenommen werden soll.
Im Unterschied zur klassischen Terminologiewissenschaft, deren Zugriff auf den genannten Untersuchungsgegenstand als in wesentlichen Punkten normativ und kaum an linguistischen Fragestellungen interessiert charakterisiert werden kann, folgt die vorliegende Arbeit dabei mit dem korpuslinguistischen Paradigma einer Forschungsmethodik, die in den letzten Jahren und Jahrzehnten an Bedeutung gewonnen und zunehmend elaboriertere Untersuchungsverfahren hervorgebracht hat. Nicht nur soll damit der Versuch unternommen werden, eine Brücke zum aktuellen Forschungsstand der Linguistik zu schlagen, vielmehr sollen auch Ansätze aufgezeigt werden, mit deren Hilfe die bisweilen kontrovers diskutierten theoretischen Konstrukte der Terminologiewissenschaft operationalisiert und – die Wichtigkeit dieses Punktes kann nicht genug betont werden – für die terminologische Praxis modelliert werden können.
Die Arbeit besteht aus sieben Kapiteln und ist wie folgt aufgebaut. Kapitel 1 leitet in das Themengebiet ein und benennt die wesentlichen Forschungsfragen. In Kapitel 2 wird der Begriff des „wissenshaltigen Kontexts“ theoriebasiert geschärft und von ähnlichen Begriffen abgegrenzt. Kapitel 3 umreißt den aktuellen Forschungsstand zum Thema. In Kapitel 4 wird die empirische Datengrundlage für die in den Kapiteln 5 (linguistische Beschreibung) und 6 (Extraktion) zu leistende Untersuchung erarbeitet. Kapitel 7 schließlich fasst die Ergebnisse der Arbeit zusammen.

This thesis is dedicated to the study of terminological concept descriptions, i.e. semantic descriptions of domain-specific concepts in different types of non-fictional texts. Terminological concept descriptions describe important attributes of concepts and relate concepts to neighbouring concepts, thus constituting an indispensable prerequisite for the transfer of conceptual knowledge. It is this quality that makes them a relevant object of scientific study. For the same reason, terminological concept descriptions or “knowledge-rich contexts” are of major importance for knowledge-based activities, such as translation, which is the field of application of this study.
While classical terminology science has approached semantic concept descriptions mainly from a normative perspective, showing only little interest for linguistic inquiry in this field, this thesis adopts a corpus-linguistic approach. It thus follows a scientific methodology that has gained ever increasing importance during the last years and decades and developed sophisticated research methods. By doing so, this thesis attempts not only to apply linguistic state-of-the-art methods to the study of knowledge-rich contexts, but also to point to ways of modelling the theoretical constructs of terminology science, which have been subjected to sometimes harsh scientific criticism, in the context of practical terminology work. The importance of this latter point cannot be over-estimated.

The thesis consists of seven chapters. Chapter 1 provides an introduction to the topic and poses the main research questions. Chapter 2 defines the notion of “knowledge-rich context” on a theoretical basis and distinguishes it from other kinds of concept descriptions. Chapter 3 summarizes the relevant state of the art. Chapter 4 provides a description of the data underlying the experiments in chapters 5 (linguistic description) and 6 (automated extraction). Chapter 7 summarizes the results.

Anne-Kathrin Schumann
geboren am 04.11.1983 in Gera

Dienstanschrift:
Universität des Saarlandes, Institut für Angewandte Sprachwissenschaften sowie Übersetzen und Dolmetschen
Campus A2.2
66123 Saarbrücken
anne.schumann@mx.uni-saarland.de, +49 681 302-2563

Ausbildung

Seit Wintersemester 2011/2012	Promotion am Zentrum für Translationswissenschaft der Universität Wien
Seit April 2011	Teilzeit-Studium Informatik (B. sc.) (FU Hagen)
2003-2009	Studium Diplom-Übersetzen an der Universität Leipzig
2002-2003	Studium der Journalistik und Kulturwissenschaft an der Universität Leipzig
2002	Abitur am Goethe-Gymnasium Gera

Relevante Berufstätigkeit

Seit April 2013	Wissenschaftliche Mitarbeiterin am Lehrstuhl Teich des Instituts für Angewandte Sprachwissenschaften sowie Übersetzen und Dolmetschen der Universität des Saarlandes im Rahmen des Exzellenz-Clusters „Multimodal Computing and Interaction“
August 2010-Februar 2013	Early Stage Researcher im Rahmen des Projekts CLARA (EU FP7) in Zusammenarbeit mit dem lettischen Unternehmen Tilde, Riga