universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Predicting Keywords in a Corpus of Field Post Letters

verfasst von / submitted by

Emilie Sitter, BA BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von / Supervisor:	Univ.-Prof. Dr. Benjamin Roth

My sincere gratitude goes to Drew Bergerson and Laura Fahnenbruck who got me excited about Hilde and Roland's letters long before writing this thesis. They have consistently provided unwavering support throughout my time as an intern and project member, and later their invaluable expertise significantly contributed to the completion of the thesis.

I would also like to thank the members of my tandem groups and Thomas Muntschick for their insightful ideas and our engaging discussions.

Abstracts

Abstract in English

The Public History project *Alltag im Krieg* publishes the correspondence of the married couple Hilde and Roland Nordhoff, who exchanged more than 2,600 letters between 1938 and 1946. Within the project, the first half of all letters have already been annotated with several out of a total of 81 thematic keywords. The first goal of the master's thesis is to automatically annotate the second half of the correspondence, based on the already annotated letters. Various text classification models are trained and compared.

In the second step, the thesis investigates the applicability of the assigned thematic keywords for a Distant Reading method. Its assumption is that in periods in which keywords are, for instance, more prevalent than average, the corresponding topics were increasingly discussed in the letters and might have had political and/or social relevance beyond the spouses' sphere. The objective is to assess how effectively readers can glean insights into the content and context of the letters based on the relevance of keywords during specific periods. After an exploratory approach, the thesis investigates five hypotheses concerning specific thematic keywords and their anticipated trends.

Examining the individual keywords provides some indications that their frequencies are related to personal, social, and political events. Unfortunately, neither the pre-existing keywords are correct with certainty, nor can a model which was trained on these noisy keywords predict highly accurate labels for the second half of the correspondence. This leads to large uncertainties in the examined keyword frequencies.

Zusammenfassung auf Deutsch

Das Public-History-Projekt *Alltag im Krieg* veröffentlicht die Korrespondenz des Ehepaars Hilde und Roland Nordhoff, das zwischen 1938 und 1946 über 2.600 Briefe austauschte. Die erste Hälfte dieser Briefe wurde bereits im Rahmen des Projekts mit jeweils mehreren von insgesamt 81 thematischen Schlagwörtern annotiert. Ziel dieser Arbeit ist zuerst, auf Grundlage der bereits annotierten Briefe die zweite Hälfte der Korrespondenz automatisch zu verschlagworten. Dazu werden verschiedene Modelle zur Textklassifizierung trainiert und verglichen.

Im zweiten Schritt untersucht die Arbeit die Anwendbarkeit der zugeordneten thematischen Schlagwörter für eine Distant-Reading-Methode. Im Mittelpunkt steht die Frage, wie gut Lesende den Inhalt der Briefe und ihren Kontext kennenlernen können, indem sie die Relevanz der einzelnen Schlagwörter in bestimmten Zeiträumen betrachten. Dafür werden Schlagwörter untersucht, die in bestimmten Perioden beispielsweise überdurchschnittlich häufig vertreten sind. Die Annahme ist, dass einige der von diesen Schlagwörtern erfassten Themen in diesen Zeiträumen in den Briefen vermehrt diskutiert wurden und auch über die persönliche Sphäre des Ehepaars hinaus politische und/oder soziale Relevanz hatten. Es wird zuerst ein explorativer Ansatz verfolgt, danach werden fünf Hypothesen über bestimmte thematische Schlagwörter und ihren erwarteten Verlauf untersucht.

Die Untersuchung der Schlagwörter gibt einzelne Hinweise darauf, dass ihre Häufigkeiten in bestimmten Zeiträumen in Zusammenhang mit persönlichen, gesellschaftlichen und politischen Ereignissen stehen können. Leider sind jedoch weder die bereits existierenden Schlagwörter, die auch als Trainingsdaten für die Klassifizierung fungieren, mit Sicherheit korrekt, noch kann ein Modell mit dieser Daten-Ausgangslage die zweite Hälfte des Briefwechsels mit hoher Genauigkeit verschlagworten, was zu großen Unschärfen in den untersuchten Schlagwortfrequenzen führt.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hilde Laube (1920—1991) and Roland Nordhoff (1907—1998) exchanged love letters from May 1938 to February 1946. In the initially formal letters, they quickly built a connection to each other that led to their wedding in 1940. Only a few weeks later, Roland was summoned for military service, and from then on, their love letters served primarily to bridge the geographical gap in their long-distance relationship. Hilde stayed with her parents in Oberfrohna, a village in rural Saxony, while Roland was stationed as a military scribe in Greece, Bulgaria, Romania, Crimea, and later Kiel, Germany, where he also had to spend several months as a prisoner of war of the British. Over the course of almost eight years, the spouses exchanged more than 2,600 letters, in which they told each other about their everyday lives, their work, their dreams, feelings, fears, and worries during the wartime years, and above all declared their profound love for each other.

Hilde and Roland's correspondence is a unique firsthand account of the everyday life of "common people" in World War II. To preserve their letters and ensure their accessibility to future generations and to make them available now to anyone interested, an international team of volunteers, interns, and students has successively been transcribing and publishing the letters since 2011. To provide a platform for the letters, Professor Andrew Bergerson (University of Missouri—Kansas City) founded the project "Trug&Schein" in collaboration with Thomas Muntschick, an affiliate of Hilde and Roland's family.

With the dedicated efforts of Andrew Bergerson and the historian Laura Fahnenbruck, in 2023 the so far published letters were successfully migrated from the meanwhile outdated and corrupted *Trug&Schein* blog to the new platform "Alltag im Krieg" ("Everyday Life in War"). Approximately half of all letters in the possession of Roland and Hilde's family have been transcribed and published to date (August 2023). I have been a member of *Trug&Schein/Alltag im Krieg* since 2020 and have thus contributed to the transcription and publication of a large number of letters, which has allowed me to gain a deep insight into Hilde and Roland's touching correspondence.

When the letters are published on the *Alltag im Krieg* website, project members also assign them thematic keywords for improved searchability. Transcribing, proofreading, and assigning the appropriate keywords to one letter is a tedious process that involves collaboration among several project members. The transcription software Transkribus provides the initial versions of all transcripts, which facilitates and shortens the necessary steps in the publication process, even though the transcripts still require meticulous proofreading. The original goal of this master's thesis was to automate the process of assigning the thematic keywords. A text classifier trained on the already labelled letters should have been used to accurately label the letters that have not yet been published.

Unfortunately, it turned out that due to several factors, the quality of the keywords predicted by the classifier is too poor to use them when publishing a letter on the *Alltag im Krieg* website. Accordingly, I had to adjust the goals of the thesis project and to shift its focus. Even without the claim to use the predicted keywords for the published letters, still some interesting research questions arose.

Primarily, the master thesis deals with a machine learning problem. The entire set of possible keywords that can be assigned to one letter consists of 81 different words. Since each letter is labelled with several out of them, assigning the keywords is both a multi-class (more than two possible classes) and multi-label (more than one assigned label) classification task. A rule-based classification model is compared to a Logistic Regression model, a linear SVM and an SVM with sigmoid kernel, and a bidirectional LSTM to identify the model that is most appropriate for assigning the correct keywords to Hilde and Roland's letters.

Due to several sources of noise in the data (imbalanced, ambiguous, and overlapping class labels as well as no real ground-truth for training, development, and test data), the best achievable performance is limited. Even the most appropriate model cannot perfectly predict the correct keywords. The focus is therefore on fine-tuning all models and identifying the best one among them, given all issues concerning the data. The best possible micro average F1-score over all classes that can be achieved on the test data is 0.34 (ranging between 0 and 1), by a stacked classifier applying majority voting of the predictions of the rule-based model, the Logistic Regression, and the SVM with sigmoid kernel. The labels that this stacked classifier predicts on all the letters that have not yet been labelled will not be used for publication, however for further analyses within the thesis. These will be the first thorough examinations in the AiK project of the keywords and the topics they convey.

The keywords serve as a starting point for a Distant Reading (Moretti 2013) method. I generate line graphs to depict the progression of all keywords for which the best model achieves an F1-score on the test data that at least surpasses the performance of a random baseline. The idea behind the graphs for the relative frequency of each keyword (the proportion of letters in a month labelled with it) is that they potentially reflect how important the corresponding topic was for Hilde and Roland in a certain period. The keywords cover personal as well as political and social topics. Accordingly, graphs for keywords related to politics and war can be expected to reflect major events of World War II.

The graphs for some of the keywords which were studied more closely demonstrate that the method does function under certain conditions. Some of the graphs appear to lack any significant pattern and exhibit no notable rise corresponding to wartime events beyond Hilde and Roland's sphere. This can be attributed to either the poor quality of the keywords (both the already existing ones and the ones assigned by the classifier) which distorts the graphs, or the fact that Hilde and Roland simply did not perceive the topic in the way it was anticipated. Under a certain idea of Hilde and Roland's perception of a topic, at the beginning hypotheses are formulated about what some graphs might reflect. But the idea might also be wrong, and for Hilde and Roland a certain topic gains or loses relevance at completely different times than anticipated. However, for instance, the graph for the keyword “Luftkrieg” has a dramatic increase that can undoubtedly be linked to the bombings of Dresden and Chemnitz in 1945, located not far from Hilde’s home village of Oberfrohna.

The thesis is structured as follows:

In Chapter 2 I specify the research questions that have just been raised here and present the hypotheses which will be examined. Chapter 3 provides background information on the correspondence of Hilde and Roland and the digitization and publication of the individual letters. It contextualizes the correspondence within the framework of the German field post and introduces the letters’ authors and the project *Trug&Schein/Alltag im Krieg*. An exploratory data analysis gives an overview of the corpus. The chapter is also dedicated to the thematic keywords with which all published letters are labelled, and which will later serve as class labels. Chapter 4 outlines the methods employed for working with the data. A reference to similar projects or analyses of similar data should make it comprehensible why precisely these methods were chosen. Due to the two distinct tasks (machine learning and interpretation of the keywords) the analysis in the chapter after is divided into two different sections. To avoid the appearance of disjointedness, the chapter embeds the overarching objective of predicting and interpreting the thematic keywords in the context of Digital History. Chapter 5 documents the practical work with the data. It describes how I pre-processed all data for the following analyses. After the pre-processing, at first, the results of the training of the various models are reported. Secondly, the complete corpus will be labelled by the most appropriate classification model. The obtained keywords for each letter then allow the creation of line graphs, which support the investigation of the frequencies of the individual keywords. In this manner, I can investigate the research questions and hypotheses posed earlier. Finally, Chapter 6 summarizes the thesis and gives an outlook on future work with the Oberfrohna correspondence.

The thesis contains several relevant passages of letters from Hilde and Roland. Unless otherwise stated, their translations from German to English are mine. Passages marked as “unpublished” cannot be found yet on the project website as of August 2023.

Chapter 2

Research Questions

2.1 Text Classification: Finding the Best Model

The first part of the practical investigation of the master's thesis focuses on training a text classifier. Through the comparison of various models aimed at predicting keywords for each letter of the Oberfrohna correspondence, the central inquiry is as follows:

Which model can predict the most accurate keywords for the Oberfrohna correspondence?

The models are confronted with both a multi-class and multi-label classification task (see Aggarwal 2015c, 169). This entails the assignment of multiple thematic keywords from a total of 81 to each document. On average, one document of the labelled data has 7.87 keywords (class labels).

Around the first half of the corpus of letters (1,309 letters) has been transcribed, proofread at least twice, and labelled in a semi-automatic manner. These letters will serve as training, development, and test data. The remaining 1,317 letters have not been labelled yet and their keywords will be assigned automatically for further analysis. To achieve this, different text classifiers will be trained, and subsequently, their outcomes will be compared. A rule-based classifier serves as a benchmark to measure which performance can be achieved without training a machine learning model. Then, Logistic Regression and a support vector machine (SVM) will be trained for the correspondence. In the last step, a deep neural network (bidirectional LSTM with CNN layer) is trained. The performance of the models is measured using the micro average F1-score, which balances precision and recall across all classes, comparable to the overall accuracy of the classifier.

In this context, there is also the question of how the nature of the data and the labels make it difficult to train a good model. The data and their labels are noisy, and the noise vastly affects the models' performance. Therefore, besides the actual training, it requires an extensive discussion of why the keywords cannot be predicted very accurately and why the approach is inherently flawed. It must be emphasized that the aim is to achieve the best possible performance, especially in comparison to the other models. Labelling the letters automatically so that they can later be published with the assigned keywords on the project's website is unfortunately not possible due to the amount of noise in data and labels. The goal is to pre-process the data appropriately and fine-tune the parameters of the models accordingly to achieve the best possible performance, no matter how good it can get with the flawed data.

2.2 Interpretation: Studying Keyword Frequencies

The second part of the analysis is dedicated to interpreting both the original and the automatically assigned keywords. The study focuses on the frequency of keywords over time, potentially uncovering trends or, conversely, observing frequency stability. Specific intriguing and frequently occurring keywords are selected for an in-depth investigation. If the letters are labelled approximately correctly with the thematic keywords, it can be assumed that the keywords and their occurrence can reveal something about the topic focuses in the course of the correspondence. If a trend or a surprisingly frequent occurrence of a keyword in a certain period can be observed, it can be attempted to relate this to real events. If connections can be made, this would mean that thematic keywords may well function to develop a Distant Reading method and reveal something about the corpus and thematic focuses in the letters.

When interpreting the keywords and their connections to real events, a high level of cautiousness is needed. Heuser and Le-Khac (2012, 30) warn researchers who work with big quantities of data that "the path from data to conclusions is tricky, lined with potential pitfalls, from misrepresentation of the data to overblown readings that stretch the data beyond what they can reasonably support." For instance, visualizing timelines of keywords in the Oberfrohna letters regarding a specific research question might reveal patterns that are, in fact, random. Challenges in interpreting the results may arise due to discontinuities in the correspondence, along with the potential issue that the keywords (both the ones assigned by the text classifier and the original ones on the labelled data) are not reliable. Moreover, the individual research questions have to be adapted to the classifier's ability to predict a keyword at all. Among the 81 classes, for 28 of them, the overall best classifier achieves an F1-score of 0 or performs worse than random on the test data. Therefore, these classes have to be excluded from further analysis. Some keywords covering cultural activities, such as "Theater" ("theatre"), "Musik" ("music"), "Oper" ("opera"), and "Film" ("film") occur less frequently, but when combined, they can form a meaningful number to formulate a hypothesis.

On a general and exploratory level, the following research questions will be investigated:

- Which **differences** in the writing habits, topic focuses, and mood (Scherstjanoi 2011, 122) between Hilde and Roland can be observed?
- There are overarching **"standard topics"** in almost all correspondences of military postal service, such as writing, love, calming attempts, and the daily military service. These topics are constant throughout all years of the correspondence (Humburg 2011, 79). Which keywords are the most stable in Hilde and Roland's correspondence?
- Can **trends** be observed in the occurrence of any keywords over time, beyond those of particular interest?
- If any trends can be observed, how do they **correlate** with societal and political events during the war?

On a more particular level, the following hypotheses were formulated to study specific keywords. The general questions above do not focus on specific keywords but pursue a more exploratory approach, aiming to identify keywords that unexpectedly reveal trends or continuity. The hypotheses, however, have a narrower focus on one specific keyword that due to prior knowledge of the correspondence can be expected to reveal interesting trends or peaks corresponding to real-world events:

- **H1:** The frequency of the keyword **"Luftkrieg"** ("air war") is directly tied to real-time events, showing an increase whenever actual attacks occur. Confirming this hypothesis would demonstrate how Hilde and Roland's private love letters can provide insights into the occurrence of air raids in World War II through the thematic keywords.
- **H2:** The frequency of keywords concerning **culture** (Kultur ("culture"), Kunst ("art"), Theater ("theatre"), Oper ("opera"), Musik ("music")) decreases over the course of the war. This hypothesis can serve to illustrate that the challenging economic circumstances and difficult living conditions at the end of the war influenced the life and priorities of especially Roland, given his identity as a musician and art enthusiast.
- **H3:** The frequency of the keyword **"Wirtschaft"** ("economy") increases towards the end of the war, as more economic and financial concerns are discussed. Considering the deteriorating economic situation in Germany during the late war years, which had a widespread impact on the population, this keyword might be able to reflect growing economic concerns of Hilde and Roland.
- **H4:** The keyword **"Kulturkontakt"** ("cultural contact") temporally increases each time Roland is newly deployed to a different location during his war service. The hypothesis should thus be consistent with Humburg's (2011, 79) notion that field post letters tend to emphasize novelty, whereas as a deployment progresses, reports about the writers' surroundings diminish in frequency since they become accustomed to them.
- **H5:** It can be observed from the letters dealing with the topic of **hygiene** that ideas of hygiene not only refer to physical cleanliness but also include racist ideas. Built upon Kipp's (2011; 2014) assumption that the concept of cleanliness was often intertwined with racial ideology during the Nazi era, this hypothesis delves into the exploration of racial ideological ideas held by Hilde and Roland.

Chapter 3

Background: The *Alltag im Krieg*—Oberfrohna corpus

3.1 German Field Post in World War II

Before diving deeply into the lives of Hilde Laube and Roland Nordhoff and their correspondence, the German field post, through which the spouses communicated between 1940 and 1945, will be introduced in general terms. According to the official statistics of the German Military Archive in Freiburg, during World War II, 31,421 billion letters were sent via the German military postal service ("Feldpost") (Buchbender 2011, 19). Each unit of the German army was assigned a unique identification number, which remained consistent even if the unit relocated (Schwender 2014, 213). Already World War I led to an extreme increase in postal services. People were confronted with sudden long-term separations and attempted to bridge them with cards, packages, and letters, which were often intended as signs of life and expressions of love (Hämmerle 2023). The main goal was usually to maintain the relationship between the deployed soldier and his family, relatives, and friends at home (Humburg 2011, 78).

Hämmerle (2023) coined the term "doing emotion" to describe the act of composing love letters during periods of separation: This practice is believed to fulfil the emotional requirements of lovers across geographical distances. When interpreting field post letters, the ones which were intended as love letters must be interpreted as such, and a love letter must be considered a genre on its own. When examining these letters, it is crucial to consider the multitude of historically evolving writing practices that influenced their composition. These traditions are intricately connected to the literary traditions of Romanticism and emphasize the significance of genuine preservation and authentic expression of emotion.

Moreover, researchers must always exercise caution with interpretations due to censorship. On the one hand, there has been external censorship. External censorship aimed to prevent the dissemination of information regarding official events that were subject to secrecy, the sending of enemy propaganda, photos under secrecy, and statements critical of the regime (Humburg 1992, 71). While the threat of severe prison sentences and death penalties for violating censorship regulations was intended to discourage such behaviour, the inspection of letters in practice remained sporadic, occurring on a random basis (Jander 2011, 441).

On the other hand, internal censorship had an impact on the letters. For instance, in many correspondences, sexuality is a topic that is subject to censorship (Humburg 2011, 82). Despite the fact that Hilde and Roland held strong religious beliefs, they discussed topics related to sexuality relatively openly. Usually, they used a smaller handwriting size, paraphrased the topic with metaphors, and embedded it within the context of a planned pregnancy. Moreover, self-censorship usually affects critical topics. For the most part, field post letters do not indicate what the soldiers knew about the murder of Jews in the occupied country (Humburg 2011, 82). Humburg (2011, 82) studied 732 letters sent via military post service, written by 25 different soldiers. Only 10 out of these 25 soldiers even mentioned Jews, and overall, the topic affects only 2% of all examined letters. In letters written later than the summer of 1941, they did not find any references to Jews at all anymore. Also Jander (2011, 445) reports on the infrequent acknowledgment of guilt, which was often not addressed earnestly and openly. Instead, when a soldier was actively involved in the murder of Jews or possessed knowledge of it, he tended to refrain from further questioning the matter. Roland was not involved. The topic of guilt, however, he did not address until the end of the war, when it could no longer be suppressed (see Chapter 5.4.4, Hypothesis 5).

The Holocaust and the Nazi war crimes are perhaps the most extreme topics demonstrating how censorship affected field post letters, making them unsuitable as objective historical sources. However, Word War II has extensively been researched and its entire course is known in minute detail. Rather than being a source for objective historical facts, field post letters serve a different purpose in their interpretation (Schwender 2011, 127; 2014, 213). Their peculiarity results from the temporal and local proximity of the writer to the reflected or omitted events, in which changes in mood and attempts of interpretation become visible, as well as from the efforts at interpretation the soldiers tried to make when they told their families about these events. The most appropriate way to interpret them is as a biographical source of a soldier's life. Any other kind of analysis has to assume questionable, false, or problematic content (Scherstjanoi 2011, 117, 122). Stader (2011, 139) compares field post letters, particularly those that were publicly available during that era, with modern-day social media and blogs. These contemporary platforms also enable people to connect, yet they are often less viable as impartial sources.

Thus, field post research always requires rigorous source criticism, understanding it against the context of subjectivity and individual sensibility, and consulting other scholarly sources (Scherstjanoi 2011, 122).

3.2 The Letters of the Oberfrohna Correspondence

This chapter provides an overview of the composition of the letters and their origin. The first known letter of the correspondence of Hilde Laube and Roland Nordhoff (pseudonyms) was written on May 4, 1938, by Hilde. On the last day of the correspondence, February 16, 1946, both Hilde and Roland wrote one letter. While Hilde wrote in modern handwriting, Roland mainly wrote his letters in German Kurrent handwriting.

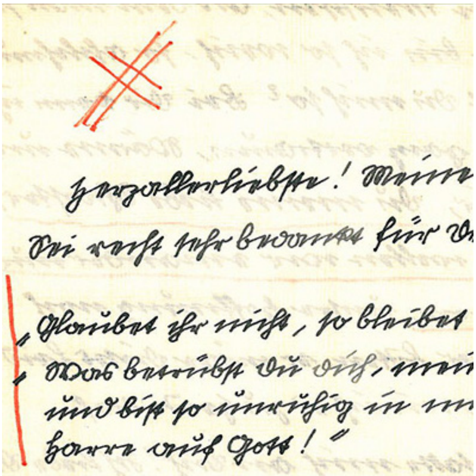
During the war's outset, they employed high-quality stationery and occasionally even wrote on fragrant paper; however, as the war progressed, the paper's quality diminished. They mainly wrote with ink (Fahnenbruck and Bergerson 2023, 2–3) and usually were scared that pencils would fade when they did not have any other option but to write with them (Fahnenbruck et al. 2023, OBF-410408-002-01)

As far as we know, at least 2,626 letters are part of the corpus. The first 1,309 of these letters have been fully transcribed, corrected, and published on the *Trug&Schein* blog (hereinafter referred to as T&S; not online anymore). One of the project leaders, Professor Andrew Stuart Bergerson, who will be introduced in Chapter 3.4, received these letters from the Nordhoff (pseudonym) family in 2011 (Bergerson and Fahnenbruck 2022, 10; Baker et al. 2020, 57). They were stored in cardboard boxes in 24 two-ring binders in the family's attic (Baker et al. 2020, 57). Since the beginning of the project, there have been gaps in the correspondence. Some of them are caused by Roland's furloughs, when the spouses were seeing each other instead of having to write letters. Sometimes, the field post letters got lost on the postal way, or Hilde did not know Roland's current address when he was travelling. (In February and March 1945, for instance, Hilde's letters arrived with a big delay and his did not arrive at all.)

Usually, the spouses attempted to exchange letters every day. When it was not possible to send the letters daily, for instance during Roland's war captivity in 1945/1946, they still tried to write letters and later sent them collectedly. Also, the spouses thoroughly archived the letters in the folders and it is unlikely that Hilde or Roland themselves lost any letters they received (Bergerson, Fahnenbruck, and Hartig 2019, 272). From time to time, the Nordhoff family themselves find another folder of letters which have not been digitized yet. With the letters contained in these folders, some gaps in the correspondence can be completed. This constant addition of missing letters indicates that 2,625 is no final number but more letters might be found.

Hilde and Roland themselves wrote the vast majority of all letters in the correspondence. Only very few letters were written by, for instance, Roland's siblings (e.g., Fahnenbruck et al. 2023, OBF-430711-003-01, unpublished). Also, the correspondence contains some additional documents, such as newspaper cuttings, photographs, or program booklets of events the sender attended. According to Baker et al. (2020, 63–64), the letters were kept by the spouses until their deaths and left untouched by the family afterwards. However, there are still additional elements in some letters, such as underlines, or red markings next to the text, as well as pages that were removed or mixed up. There is no complete certainty when or by whom they were caused. Different kinds of ink indicate that the writer of the letter did not create them during writing. Some project members and Hilde and Roland's family suspect that Roland went through all the letters again after Hilde's death, sorting them and marking the passages that were important to him (see Figure 3.1). Despite these elements, it was decided in the T&S project that authority as primary source should be ascribed to the letters (Baker et al. 2020, 64).

Instead of using the spouses' actual first and last names, full pseudonyms are employed in the transcribed letters. For any other mentioned people, their original first name is kept but their privacy is protected by using only the initial of their last name. When the letters were published on the T&S blog, originally also smaller locations (such as Oberfrohna, Roland and Hilde's home village) were abbreviated by initials, while larger locations (such as Chemnitz) were expected to not give


Herzallerliebste! Mein

„Glaubet ihr nicht, so bleibet
„Was betrübst du dich, mein
und bist so unruhig in m
harre auf Gott!"

Figure 3.1: Example of letters with markings (Fahnenbruck et al. 2023, OBF-390905-001-01)

enough clues about the spouses' identity. Since the correspondence is hosted on the AiK website, anonymizations of places have been reversed because it can be assumed that after around 80 years, the place names of towns and land sights alone are not enough information to reveal anyone's identities.

3.3 Introducing Hilde and Roland Nordhoff

The centre of the correspondence are its writers, Roland and Hilde Nordhoff (unmarried Laube, all pseudonyms). Hilde was born in 1920 in Oberfrohna, a village in rural Saxony. Roland was born in Kamenz, a Saxonian town, in 1907, which makes him 13 years older (Bergerson, Fahnenbruck, and Hartig 2019, 265). Hilde did not have any siblings, Roland had two brothers (Hellmuth and Siegfried) (Fahnenbruck et al. 2023). Before the war, Roland originally worked as a teacher and Hilde in a textile factory. Both of them were "Aryans" by Nazi law and faithful Protestant Christians (Fahnenbruck et al. 2023). They met in the choir of the local Protestant church (Bergerson, Fahnenbruck, and Hartig 2019, 265). In May 1938, Hilde wrote her first letter to Roland, starting a regular correspondence two years before his conscription into the military, in which she asked him for a conversation ("Aussprache") (Fahnenbruck et al. 2023, OBF-380504-002-01). From then on, their relationship started to develop in secret. Only in 1939, they began to tell their families and friends about it. While Hilde was still living with her parents (often called "Mutsch" and "Pappsch" in her letters) in Oberfrohna, Roland was transferred to a new position at a school in Lichtenhain (around 130 kilometres of distance). They married on July 13, 1940, when Hilde was 20. On August 27, 1940, Roland was enlisted into basic training for military service. After only six weeks of marriage, they were separated spatially again and their correspondence continued (Fahnenbruck et al. 2023).

Roland's first deployment was in Barklesby, a town in Schleswig-Holstein north of Kiel. From then on, he served as a military scribe. He described his new job to his wife like this:

Herzlieb, ich bin jetzt nicht mehr so unter der namenlosen Menge — bin mehr für mich — brauche nicht jeden Dienst mitzutun — meine Dienststelle ist die Schreibstube, ich bin jetzt, wenn auch nur ein Rädchen, in der Leitung des Getriebes — gehöre jetzt mit zum Gehirn dieser Stellung. Dieses Bewußtsein läßt sich mich in dieser Enge freier atmen. Für mancherlei entscheidende und begehrenswerte Dinge sitze ich jetzt an der Quelle. Nicht, daß ich diese Quelle mißbrauchen will. Aber ich kann ein wenig rechnen im voraus [sic] und kenne besser Stellen und Wege zu diesen Dingen: Urlaubsscheine, Wehrmachtfahrscheine, Dienstpläne, Urlaubslisten usw.

Sweetheart, I am now no longer as much among the nameless crowd—am more for myself—do not need to do every service—my office is the writing room, I am now, even if only a cog, in the management of the gear—now belong to the brain of this position. This awareness allows me to breathe more freely in this confinement. For many decisive and desirable things, I am now sitting at the source. Not that I want to abuse this source. But I can calculate a little in advance and know better places and ways to these things: Vacation tickets, Wehrmacht tickets, duty rosters, vacation lists, etc.
(Fahnenbruck et al. 2023, OBF-401102-001-01)

In March 1941, Roland was relocated to Bulgaria (Fahnenbruck et al. 2023, OBF-410326-001-01). In April 1941, he was moved to Thessaloniki, Greece (Fahnenbruck et al. 2023, OBF-410430-001-01, unpublished), and in October 1942 to Bucharest, Romania (Fahnenbruck et al. 2023, OBF-421018-001-02). He was part of the military mission sent by Hitler to Romania in 1940. The goals of the military alliance between the German Reich and Romania were for the German Reich to secure Romanian oil reserves as well as to have a starting point for an attack on the Soviet Union. Romania, on the other hand, hoped that the stationed German soldiers would deter the Soviet Union from an attack (Panthöfer 2015). Between January 11, 1944 and May 1944, Roland was on the island of Crimea. On May 13, 1944, he reached the Romanian mainland in Constanța by ferry, having left Crimea a few days before, but there are no letters available of the exact day of his departure (Fahnenbruck et al. 2023, OBF-440110-001-01, unpublished; 2023, OBF-440514-001-01, unpublished). For several weeks during his deployment on Crimea, there are no letters because it might have been impossible to write letters or because postal services were disrupted. Hilde even suspected that Roland was dead.

In April 1944, the advancing Soviet troops in Romania could no longer be held back. In Romania, the government was overthrown, the new government broke off diplomatic relations with the German Reich, and on August 24, 1944, Hitler announced the bombing of Bucharest (Panthöfer 2015). Roland's last letter from Constanța, where he had stayed since his return from Crimea, is full of uncertainty and is dated August 26 (Fahnenbruck et al. 2023, OBF-440825-001-01, unpublished). The next available letter, dated August 30, is from Niš, Serbia, on the way to Belgrade (Fahnenbruck et al. 2023, OBF-440830-001-01, unpublished). After that, Roland was allowed to go back to Germany, was able to spend time with Hilde in transit, and was finally transferred to Kiel in October 1944. In Kiel, for the first time during his military service, Hilde was also able to visit him personally. He had to stay in British war captivity until February 1946 and finally returned home in the night of February 17, 1946.

During the entire time of their separation, Hilde and Roland tried to maintain the normalcy of their marriage by shared practices of showing love, such as placing letters or photos of themselves in their rooms, listening to the radio together, picking flowers for each other, drying and then sending them, or, in the case of Hilde, sharing the letters with her family and sending their greetings to Roland (Bergerson, Fahnenbruck, and Hartig 2019, 270). They crafted the letters in a manner that they resembled genuine conversations. In his first letter from basic training, Roland wrote:

Nun zunächst meine Feldpostnummer, damit unsre Verbindung in Gang kommt, damit wir uns die Hände reichen können und miteinandersprechen, als wären wir umeinander.

Now first of all my field post number, so that our connection gets going, so that we can join hands and talk to each other as if we were around each other.
(Fahnenbruck et al. 2023, OBF-400830-001-01)

They metaphorically equated the act of writing or reading letters to the experience of physically being in each other's presence with phrases like "Du bist heute wieder so lieb, so lieb zu mir gekommen" ("You came to me again today so, so sweetly") (Fahnenbruck et al. 2023, OBF-421101-002-01) or, talking about tomorrow's plans, "Und am Nachmittag sitze ich dann bei Dir und plausch' mit Dir." ("And then in the afternoon I sit and chat with you.") (Fahnenbruck et al. 2023, OBF-421127-002-01).

Shared rituals, talking to each other in letters, and sending parcels with gifts and supplies to maintain normalcy in the relationship during their separation were not unique to Hilde and Roland but were a common practice of many couples communicating via military postal service. The correspondence helped to hold on to the relationship and was also a constant reminder for the spouses at home to stay faithful (Humburg 2011, 79). Since the letters and everything that could be sent by field post, as well as sometimes telegrams, were the only way for soldiers to communicate with their spouses and families at home, they were of particular relevance in their lives (Schwender 2011, 127). Schwender (2011, 127) writes about how this made fixed everyday communication and the maintenance of family networks the main task of field post letters, while reporting actual events at the front was either a side issue or in general neglected.

One important criterion for war events that soldiers did report in field post letters was their novelty. Typically, at the commencement of a deployment, there was a tendency to produce disparaging accounts of the local people and culture. However, as the deployment progressed, a growing sense of detachment emerged, leading soldiers to decrease their discussions regarding the host country or their duties (Humburg 2011, 79). Roland did report events at the front with a "matter-of-fact tone", making an effort to delineate his private life from his military service (Bergerson, Fahnenbruck, and Hartig 2019, 259).

According to Steele (2011, 273), letters from home hold equal significance to those from the front when considering them as a biographical source that provides insights into perceptions and emotions. They examined the role of the wives at home in field post correspondences and noted a shifting priority among these women, where communicating with their husbands takes precedence over the ongoing progression of the war. Apart from working on the relationship and a way of talking to their husbands, for many wives writing letters served an additional purpose: they could ask their husbands for opinions through the letters and the men retained a degree of influence over household decisions, a responsibility that had otherwise rested solely with the women following the men's deployments (Steele 2011, 277). The following letter is an example of Roland giving advice concerning the household:

Frag doch mal – ganz auf eigne Faust zunächst – was ein Gas- oder ein elektrischer Ofen kostet, weißt, so wie wir in Kamenz im Bad stehen haben. Und wenn es noch welche gibt – nicht zu teuer – dann greif zu – frieren wollt ihr und wollen wir doch nicht!! [. . .] und wenn solch Ofen nicht über 100 kostet, dann mach das Geschäft fest – hörst mich, Herzlieb?!!!

Ask—on your own first—what a gas or an electric stove cost, you know, as we have in Kamenz in the bathroom. And if there are still some—not too expensive—then buy it—neither you or us do want to freeze!!! [. . .] and if such a stove does not cost more than 100 Reichsmark, then make the deal firm—do you hear me, Sweetheart?!!!
(Fahnenbruck et al. 2023, OBF-410110-001-01)

Hilde and Roland's letters are interspersed with prayers and blessings, which makes belief one of the main motifs in their letters and "Glaube" ("belief") one of the most frequent thematic keywords of the corpus. The following example was taken from one out of a large number of letters with similar prayers to God: "Gottvater, Du, im Himmel! Hör unser Bitten! Segne unsre Wege, unser Wollen! Hilf uns! Amen." ("God the Father, You, in heaven! Hear our plea! Bless our ways, our will! Help us! Amen.") (Fahnenbruck et al. 2023, OBF-440320-001-01, unpublished).

Just like God, also "Führer" Adolf Hitler became an important part of their lives (Bergerson 2018, 236). Although Hilde and Roland rarely mentioned Hitler literally (Roland 11 times, Hilde 7 times in the entire corpus[1]), during Roland's enlistment they began to discuss themes like their duties and their willingness to make sacrifices for the Führer and their country. To best integrate in the *Volksgemeinschaft*,[2] Hilde aimed to conform to Nazi standards by starting a family

[1]Since the corpus contains not yet corrected HTR transcripts which could contain spelling errors in the word "Hitler", this figure might not be exactly correct.

[2]The term "Volksgemeinschaft" ("national community") was introduced by the Nazis in an

with many children early on (Bergerson 2018, 236). Roland and she discussed this topic frequently, usually in reference to God and in smaller handwriting compared to the rest of the letter. In the following letter from June 1941, Roland talked about their desire for children as well as a deeper, spiritual meaning of a family:

Weißt Du, wie ein feiner leiser Wink vom geliebten Weib das Mannerli ganz selig machen kann? Du!!!!! !!!!! !!! Mein Herzlieb hat den Sinn vom Familienbaum gewiß ganz tief begriffen und sich gemerkt, ich kenn es doch darin! Mutter – Lebensbaum – Kinder, oh Du!!! so viel!!! – wer ist denn grade dran, am Lebensbaum zu blühen und zu fruchten?

Do you know how a subtle, gentle hint from the beloved wife can make the Mannerli completely blissful? You!!!!! !!!!! !!! My dear heart has certainly deeply understood the meaning of the family tree and has remembered it, I do know this about her! Mother—tree of life—children, oh you!!! so many!!!—who is just about to blossom and to fruit at the tree of life?
(Fahnenbruck et al. 2023, OBF-410617-001-01, unpublished)

While initially they planned to wait until Roland's return before having children, later letters suggest that they did try to conceive during Roland's furloughs:

Ach, bist Du auch gewiß nicht mehr traurig darum, daß das Kalendermannerli [Menstruation] wiedergekommen ist? [. . .] Es liegt nicht an unserem Willen allein! [. . .] Ich glaube doch, daß zu unserem Wollen nichts mehr fehlt als das Wollen Gottes. Und darum kann ich doch auch nicht traurig sein, und ungeduldig.

Oh, are you also certainly no longer sad about the fact that the Kalendermannerli [menstruation] has come back? [. . .] It is not up to our will alone! [. . .] I believe that nothing more is missing to our will than the will of God. And therefore I cannot be sad and impatient.
(Fahnenbruck et al. 2023, OBF-421109-001-01)

However, in the last months of the war, they took back their early desire for children:

Jeden Tag danke ich es dem Herrn, daß er uns in diese Zeit kein Kindlein bescheren wird! Es ist ein rechter Segen. Wo nun die Ernährungslage immer kritischer wird.

Every day I thank the Lord that He will not bring us a child in this time! It is a real blessing. Now that the food situation is becoming more and more critical.
(Fahnenbruck et al. 2023, OBF-450205-002-01, unpublished)

attempt to establish a community of people in the Third Reich closing all political and social rifts left by the Weimar Republic. It claimed to be egalitarian, neglecting all differences of income, profession and education of its members, and that Germans were one unit in solidarity. It was also racially based and played a central role in establishing the totalitarian system of rule (Scriba 2014).

In the end, their first son (of three children) was conceived soon after Roland's return from war captivity on February 17, 1946, and born around nine months later, on November 26, 1946. Bergerson (2018, 236) assumes that Hilde's early marriage and desire to have children might have been motivated by a potential exemption from Nazi labour service for married women or women with children. However, the regulations regarding who was eligible for exemption were unclear and even varied locally.

Like other enlisted soldiers, Roland adhered to certain ideologies associated with the Nazi regime, at least to some degree. He read an antisemitic paper and in the letters to Hilde, he reflected on the ideological goals of the war and tried to find the political and spiritual significance of his military services (Bergerson, Fahnenbruck, and Hartig 2019, 263). In the last years of the war, however, a rejection of National Socialist ideology and a condemnation of the atrocities can be observed more and more strongly in Roland's letters (see Chapter 5.4.4).

During the entire time of their separation, Hilde and Roland maintained continuous correspondence. On rare occasions, when one of them couldn't find time to write for an entire day, they had a common practice of writing one letter the morning and another one in the evening of the following day. Steele (2011, 280) observed in their study about women's everyday life in World War II in Vienna that field post letters would become less frequent or were interrupted completely when they were delayed for weeks or failed to arrive, as it "took much more energy to carry on writing into silence and uncertainty and somehow keep a workable future alive." However, this was not the case for Hilde and Roland. Particularly in some of Roland's letters from the end of 1945 and the beginning of 1946, during his war captivity, it is evident that he never sent them but instead gave them to Hilde after his return. He wrote them in a notebook one after another for weeks, usually starting a new letter without starting a new page.

This brief overview of Hilde and Roland's life between 1938 and 1946 could already introduce some of the most important themes in the correspondence. Their corresponding keywords in the labelled letters are, for instance, about belief ("Glaube"), family ("Familie"), national socialism ("Nationalsozialismus"), values ("Werte"), and war ("Kriegsschauplatz", "Kriegsverlauf", "Kriegsfolgen", "Luftkrieg"). Moreover, a recurring theme which is frequently discussed in the letters is the relationship itself ("Paarbeziehung"), along with sexuality ("Sexualität") and expectations for the future ("Zukunft"), and, most importantly, feelings ("Gefühle").

Expressions of mutual love can be found in nearly every letter. They are represented by the keyword "feelings", which is problematic due to its extent and how it was created (which will be explained later in Chapter 3.7), and does not necessarily refer only to positive feelings and love. However, love was naturally the most important and most present feeling in the love letters. To convey their feelings, Hilde and Roland developed an idiolect. They used specific salutations and created a broad range of nicknames (Bergerson, Fahnenbruck, and Hartig 2019, 257), like "Mannerli", "Herzallerliebster, "Schätzelein" for Roland, "Weiblein", "Herzensweibe", "Geliebte" for Hilde, and "Herzelein" for both of them. They often addressed themselves in the third person using these nicknames—usually

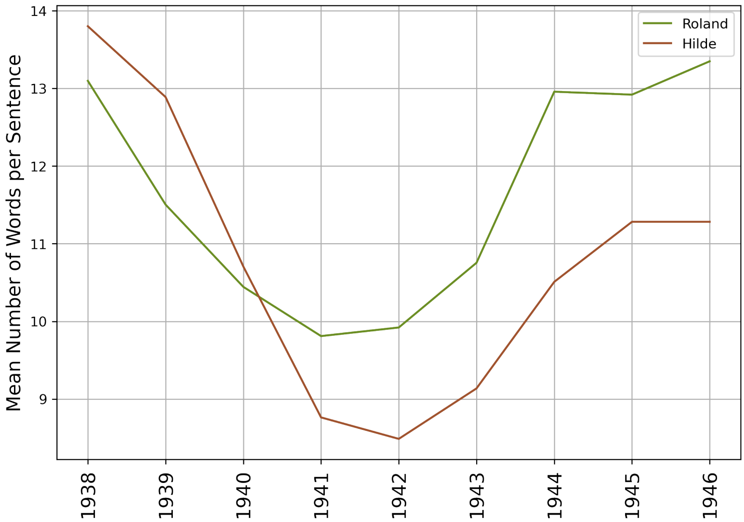


Figure 3.2: Average Sentence Length

“Mannerli” and “Weiblein” (diminutives for husband and wife)—such as: “Weißt, wie es da Deinem Mannerli zumute war?” (“Do you know how your Mannerli felt?”) (Fahnenbruck et al. 2023, OBF-410111-001-01). For text pre-processing and training the models, due to their high frequency, the nicknames can be added to a stop word list (see Appendix B.1).

Using diminutives was not limited to nicknames but extended to various contexts, for instance, animals, landmarks, or anything they considered nice or beautiful. Hilde and Roland primarily used the German diminutive suffixes “-lein”, which is especially noticeable when they referred to babies as “Kindlein” (“child” + diminutive suffix), and -(e)l, as in “Kirchl” (“church” + diminutive suffix), or “Strümpel” (“socks” + diminutive suffix). Moreover, they employed Saxonian dialectal and colloquial terms, such as “Drasch” (East Middle German for *stress*). Roland tended to use poetic and sometimes archaic language, for instance, “Antlitz” instead of “Gesicht” (“face”) (see Fahnenbruck and Bergerson 2023, 5–6). Additionally, the size of their handwriting carried meaning. A larger size often signified enthusiasm, excitement, and affection, while they used smaller letters for discussions of intimate matters exclusively shared between them, such as sexuality or their aspirations for children (compared to “whispering” by Fahnenbruck and Bergerson (2023, 4)).

Inspecting the average sentence length in their letter over the years (see Figure 3.2) makes not only visible how similar their writing style was but also how synchronised its change was. While the decrease in the average number of words per sentence in the first few years of their relationship could be attributed to deepening trust in each other and therefore a successively less formal writing style, it remains challenging to ascertain the reasons behind the subsequent increase in the sentence length of both of them and particularly in Roland's after 1941/1942. It has to be kept in mind that the transcripts from September 1943 to the end of the correspondence in 1946 have not been proofread yet and therefore might distort the graph. Punctuation marks are frequently either completely overlooked (e.g., in the case of periods) by Transkribus or incorrectly transcribed (e.g., as a colon instead of an exclamation point). In Figure 3.2, the average sentence length in the uncorrected letters may therefore appear longer than it actually is. However, the gradual and continuous increase of the graph from 1943 onwards speaks against this assumption. If the increase in the average sentence length were only due to transcription errors, there would have to be one abrupt increase of the average sentence length in 1944 at the transition from proofread to raw transcripts.

3.4 From *Trug&Schein* to *Alltag im Krieg*

The Oberfrohna correspondence is accessible on the *Alltag im Krieg* (hereinafter referred to as "AiK") project website as one of three correspondences between married couples during World War II (Fahnenbruck et al. 2023). However, the initial intention for the digitization and publishing of the letters was established within the context of a public humanities project titled "Trug&Schein: Ein Briefwechsel." This chapter is dedicated to the establishment of the T&S project and its evolution into AiK in 2022/2023.

The first project title "Trug&Schein" ("swindles and shams") originates from a letter from 1938, in which Roland described his impression of peoples' behaviour in contemporary society (Baker et al. 2020, 55):

Wir leben in einer schweren Zeit. Trug und Schein verhüllen die Wahrheit, alle Menschen t[ra]gen irgendeine Maske, rohe Lust und Begierde spielen sich überall frech auf, und es ist ein Glück, eine Gnade, wenn man gerade und unverbogen bleibt, wenn man den Versuchungen nicht erliegt und sich den Glauben und die Sehnsucht nach dem Guten, Echten und Edlen herüberrettet.

We live in dark times. Swindles and shams cloak the truth. Everyone wears some kind of mask. Raw lust and cupidity show off everywhere, and it is a stroke of luck, a blessing, if one can remain straight and unbowed, if one does not succumb to the temptation and can salvage from it one's faith and yearning for what is good, true, and noble.
(Fahnenbruck et al. 2023, OBF-380516-001-01, translated by Baker et al. 2020, 55)

Professor Andrew Stuart Bergerson (teaching at the University of Missouri-Kansas City), a historian whose interests lie in ethnographic and oral history, digital and public humanities, and the history of everyday life especially in Germany (Curators of the University of Missouri 2023), learned about the correspondence first in 2009 from Thomas Muntschick, who is also an historian, as well as an affiliate to the Nordhoff family (Bergerson and Fahnenbruck 2022, 10). Together, they launched the project. Originally, the main intent of the project was to digitize the letters to make them accessible to a broad audience on the web (Baker et al. 2020, 63). Within the next few years, the focus changed to the participatory, intergenerational, intermedial, and transnational approach of T&S. By 2011, a full project concept had been developed by Andrew Bergerson and Thomas Muntschick in agreement with the Nordhoff family (Bergerson and Fahnenbruck 2022, 10), whose condition was to give as few indications of their identity as possible and to keep them anonymous (Bergerson and Fahnenbruck 2022, 1). Rather than simply documenting the Nazi past, from then on, the goal of the project was to encourage the public to engage in reflections and critical discussions about the meaning of the past for the present (Bergerson, Muntschick, and Schwartz n.d.).

Cohen and Rosenzweig (2005, 144) mention in their guide to digital history that "building your audience by supporting and connecting with communities is not only the least expensive and most effective way of promoting your site, but also the one that most likely supports your larger social goals." T&S's goal was a slowed down, decelerated ("entschleunigt") (Bergerson and Fahnenbruck 2022, 2) approach to contemporary history, based on crowdsourcing and the participation and input of an interdisciplinary and international team of students, scholars, and citizen scientists from the United States and several European countries. These project members transcribed the letters and/or corrected transcription errors, contextualized the letters' content with historical background information, and published them online (Bergerson, Muntschick, and Schwartz n.d.; Baker et al. 2020, 45–55). Each of the participants was explicitly invited to bring in their perspective, questions, and memories, from either their own scientific or social background (Baker et al. 2020, 63; Bergerson and Fahnenbruck 2022, 2).

Community building and creating networks of participants are crucial to achieve this goal. Therefore, on the one hand all participants are in close interchange with the project leaders (Andrew Bergerson and since 2015 also Laura Fahnenbruck, a German historian working in Groningen/Netherlands). On the other hand, collaboration between participants is constantly encouraged. In 2012, the working group "Freie Altenarbeit Göttingen" (FAG) was founded to enable collaboration of elder people who were able to read the German Kurrent handwriting and younger project members with better computer skills. Before the introduction of the transcription software Transkribus to the project in 2018, the FAG manually transcribed most of the letters from 1941 and 1942 (Bergerson and Fahnenbruck 2022, 12).

Aside from the group in Göttingen, since 2011 there have been "tandem" groups of international students and citizens working together to transcribe, discuss, and contextualize the letters (either remotely or in person) (Baker et al. 2020, 64). For instance, as the author of this thesis and a longtime project member, I have been part of a tandem group since February 2022, consisting of one German and one US student, a university teacher from the US, a senior citizen from Germany, and me (Austrian student). In each of the two-weekly meetings, we select one or more letters in their raw version after the initial transcription by Transkribus, read it together and correct transcription errors while reading. During and after reading, we discuss the letter's content, try to contextualize it with prior knowledge, do some further research on unclear passages, and possibly find references to the present. The outcomes of our discussions are published as a comment when the processed letter gets uploaded on the website. The comments offer the opportunity for further discussion.

Cohen and Rosenzweig (2005, 4–6) point out the advantage of digital media for historians beyond storage capacity and easier accessibility: "The web [...] has given a much louder and more public voice to amateur historians." They also mention the importance of discussion boards (Cohen and Rosenzweig 2005, 152). Summarizing and publishing the discussions' outcomes in the former T&S project and now in the AiK project makes sure that they are preserved. Moreover, the documented discussions can be valuable input for researchers who want to work with the letters later or for any other interested person who reads through the letters. Vivid discussions in online history projects also make the websites look active and lived-in and might encourage more people (in the case of T&S/AIK other visitors of the blog, apart from the regular participants) to think about the letters and comment them (Cohen and Rosenzweig 2005, 151). Due to the participants publishing their comments, T&S/AiK does not run the risk of not getting any participation. One potential risk, however, are controversial discussions which could arise (Cohen and Rosenzweig 2005, 152). To prevent the risk of publishing comments in T&S/AiK with either controversial content or, for instance, violating the privacy/anonymization policy of the project, every comment has to be reviewed and approved by a designated project member.

Several other projects developed originating from T&S. Since 2013, there has been a monthly radio broadcast of the letters on the non-commercial radio station Radio Tonkuhle (Bergerson and Fahnenbruck 2022, 11). Andrew Bergerson, K. Scott Baker, and Deborah Parker adapted the letters into a theatre play titled "Love in the Age of Hitler: A Courtship in Letters, 1938–1940" (and its German translation "Eine Liebe in Briefen, 1938–1940"). Lena Faecks developed a video game based on the letters (Bergerson and Fahnenbruck 2022, 11–13). Moreover, learning modules to use the correspondence for teaching both history and the German language have been developed (Bergerson and Fahnenbruck 2022, 1).

Between 2013 and 2023, the T&S project was hosted on a WordPress blog. In the end, excessively long loading times and functionality issues made the website hardly usable. In April 2023, the T&S blog went offline. The correspondence has been transferred to a new website, *Alltag im Krieg*. It is built with Drupal[3] and hosts two more field post correspondences from World War II. To keep them apart, the correspondences are now called after their place of origin. Hilde and Roland's correspondence can be found as "Oberfrohna" now, named after Hilde's home village (Fahnenbruck et al. 2023). Since transferring the letters which have already been published on T&S to the AiK website requires additional processing again, not all letters have been transferred yet. In August 2023, 904 letters can be found on AiK, with additional letters pending transfer.

With the transfer to AiK and thus the emergence of a new project, the T&S project can be considered as completed. In contrast to the participatory approach of T&S and the goal to encourage everyone to engage deeply with contemporary history, AiK's focus is rather on the simple digitization and preservation of the

[3] https://www.drupal.org/

three correspondences it comprises (Fahnenbruck 2023). This shift in the project goals means that participants no longer need to provide extensive contextualization while transcribing and blogging the letters. The tandem groups, originally formed as part of T&S, still perpetuate and publish their discussion outcomes along with the letters on the website. The new approach now promises a faster publication process and provides researchers and interested public an opportunity to work not only with the correspondence of Roland and Hilde but also to compare it with the two other correspondences published on the AiK website.

3.5 The Digitization and Publication Process of the Correspondence

For the digitization and publication of the handwritten letters, which will be presented in this chapter, AiK follows a crowdsourced approach. As the former complete project title "Trug&Schein: Ein Briefwechsel. Eine kritische Begegnung mit dem Alltag des Zweiten Weltkriegs – Schreib mit!" ("Trug&Schein: An Exchange of Letters. A critical encounter with the everyday life of World War II—Write along!") suggested, T&S focused on the participatory aspect of encouraging project members to critically analyse, contextualize, and reflect on the letters with reference to the presence and their own everyday life (Bergerson and Fahnenbruck 2022, 2). In contrast, AiK adopts a more traditional form of crowdsourcing history, involving citizen scientists primarily in the digitization and publication processes of the letters.

In its core definition, crowdsourcing refers to "large groups of nonexpert, low-paid workers or volunteers performing various well-defined tasks" (Suissa, Elmalech, and Zhitomirsky-Geffet 2021, 271). In the context of crowdsourcing cultural heritage, the term is problematic due to its association with work/labour (Owens 2014, 270). Owens (2014, 270) advocates for inviting the public to contribute to projects in this field instead of thinking of crowdsourced projects as "outsourcing" labour. In AiK, most project members indeed do not consider their contributions as "work". The closest activities resembling "work" involve internships for students in a field in the humanities, which usually are not monetarily rewarded, but with an internship certificate/a letter of reference. Additionally, short-term contracts are occasionally offered for urgent tasks if research funding is available. Apart from that, for most project members, their primary motivations for contributing are their interest in the content and the enjoyment they derive from editing and discussing the letters in a group setting. They consider it as a hobby. This makes AiK more valuable for all involved rather than just the project benefiting from the "wisdom of the crowd" (Opryshko and Nazarovets 2021, 200). As Owens (2014, 279) points out as one of the main advantages of crowdsourcing digital heritage by citizen scientists, the project members themselves can become authors of the historical record and public memory.

The most frequent use cases of crowdsourcing in the (digital) humanities and the field of cultural heritage are the provision of historical artefacts (texts, images, audio, and video) by image and text collection, interviews, transcription, and annotation to make them available and searchable for the public on the web (Smolarski, Carius, and Plaul 2022, 306–7; Suissa, Elmalech, and Zhitomirsky-Geffet 2021, 268, 272). The increased use of digital tools and digital media facilitates the participation of citizen scientists in research projects (Smolarski, Carius, and Plaul 2022, 305). A project like T&S/AiK would have been impossible without digital infrastructures. Even the communication between the international project team, located in Germany, Austria and the United States, and the project leaders (Laura Fahnenbruck in the Netherlands and Andrew Bergerson in Kansas City, Missouri), is only possible due to new digital media (e-mail, Zoom, and also the messenger service Signal). Like this, T&S/AiK demonstrates how new ways of communication enable collaboration among institutions and individuals in different locations who do not have formal affiliations with one another (van Hyning 2019, 4).

The work on the historical source itself is entirely based on digital technologies. The handwritten letters were scanned in 2011 by a project member. Before employing a HTR (Handwritten Text Recognition) transcription software, the letters were transcribed manually using Zotero. Since 2018, the tool Transkribus has been used (Bergerson and Fahnenbruck 2022, 10–13). Transkribus is both a software and a platform which allows researchers, archivists, and scholars from universities all around the world to transcribe handwritten and printed texts through HTR and OCR (Optical Character Recognition) (Muehlberger et al. 2019, 955). Transkribus has been developed and was launched in 2015 by the READ (Recognition and Enrichment of Archival Documents) initiative, headquartered at the University of Innsbruck (Muehlberger et al. 2019, 957). The general workflow of transcribing texts involves the layout analysis of the scanned pages since the transcription follows horizontal lines, the actual transcription using deep neural networks (Muehlberger et al. 2019, 958), and the manual correction of potential transcription errors. For a specific handwriting, users can train their own HTR model on a minimum of around 25 transcribed ground-truth pages (or 5,000 transcribed words) (Muehlberger et al. 2019, 959–61). In the T&S/AiK project, one HTR model has been trained for Hilde and one for Roland. Since there were already a lot of ground-truth data, for each model in total around 70,000 words were used for training and validation (split 90:10). Until 2022, all available letters of the Oberfrohna correspondence written between 1942 and 1946 were transcribed month by month with Transkribus.

After the automatic transcription of the letters of a given month, they are further processed by the project members. Besides the interns and members with short-term contracts, AiK members include individuals from various backgrounds, such as members of the already introduced working group "Freie Altenarbeit Göttingen", former interns, students in a related field, or people of the interested public. They become aware of the project through various channels, such as acquaintances, affiliated university institutes, internet research on related topics, or media coverage that sparked their interest. The group of members experiences constant changes, for instance, some interns or students do not further contribute to the project when they finish their internship/their research. However, many long-term members have been part of the project for years. Two of the longest project members have been part of T&S/AiK since 2014. Owens (2014, 270) points out how any interested person can participate in projects that crowdsource cultural heritage, but usually a small, dedicated community establishes. Similarly, van Hyning (2019, 8) describes a "long-term and potentially open-ended commitment" for a dedicated community. The participants are neither strictly professionals (in AiK, many members do have a background in any field of the humanities, but some do not and are just interested) nor volunteers in a literal sense. Rather, they are "amateurs in the truest and best possible sense of the term", where the term "amateur" derives from French "lover of" (Owens 2014, 271). In an analysis of the psychology of their motivation, Owens (2014, 276–77) identified various non-financial benefits that people can achieve from participation in citizen science projects, such as an improved self-image, a meaning to their life by making a meaningful contribution to the public good, and connecting with like-minded people. When engaging in crowdsourcing cultural heritage, incorporating participants as co-authors within a publication adds additional value to their contributions and imbues their efforts with significance (Owens 2014, 271). On the AiK website, all project members' names (current and former) are listed.[4]

The first step when working with the plain HTR transcripts is called "crosschecking", which primarily entails correcting HTR errors. Additionally, crosscheckers have the option to add annotations as needed. These following annotations, which are documented way more extensively and should only be introduced briefly here, are possible:

- Larger or smaller font size compared to the normal one
- Super- or subscript (for example, used when Hilde and Roland accidentally omitted a letter or word and added it later)
- Anonymizations of names (when exporting the text from Transkribus, only the first letter of an anonymized word is kept)
- Unclear letters/words
- New paragraphs (larger gaps in one line are also interpreted as new paragraphs if the topic changes)
- *SIC* for obvious spelling mistakes (idiolectal or archaic spellings, such as "garnicht" ("not at all") without space are not annotated)
- Hyperlinks to online dictionary entries for abbreviations and dialectal or idiolectal terms that most readers would not understand
- Anything else is annotated as "other", for instance, to indicate missing characters due to a hole or a stain on the paper

After crosschecking, the letters are reviewed again by other people, usually more experienced project members ("editors"). They double-check the transcripts and add (more) annotations.

Continuous proofreading and correction of transcripts contribute to the improvement of the Transkribus models by providing more training data. The

[4] https://alltag-im-krieg.de/ueber-uns

underlying neural networks continuously learn from each piece of ground-truth data they receive, resulting in an improved performance the more HTR transcripts are manually corrected (Muehlberger et al. 2019, 966–67). With the models implemented in 2019, a CER (character error rate; Transkribus scores its performance in terms of the proportion of incorrectly transcribed characters) of below 5% can be expected (Muehlberger et al. 2019, 962). As of August 2023, the CER for Hilde's model stands at 4.0%, while Roland's model has a CER of 1.4% on the validation data. To further improve the results of the HTR, users can provide dictionaries or vocabulary lists (Muehlberger et al. 2019, 961). Such lists have been provided for the transcription models of Hilde and Roland as well.

Once at least two people have corrected a HTR transcript, the letters can be exported from Transkribus and published on the AiK website. For now (by August 2023), the letters until approximately the beginning of 1944 have been crosschecked. Since all exports needed to be prepared for my master's thesis considerably before the completion of the crosschecking process for all the letters was within reach, I exported an initial, preliminary version of the entire correspondence in December 2022, albeit with some degree of noise (the export and related issues will be discussed in Chapter 5.1). Usually, Eric Bergerson handles the Transkribus export, who has developed a Python script to transform letters and their metadata into XML files designed for the Drupal import. This export occurs after crosscheckers and editors proofread the letters. After the transformation, the letters can be blogged on the website. The blogging guidelines in T&S were slightly different from the current guidelines in AiK. The letters do not require as much contextualization anymore (for instance, project members should not insert images/photos of events mentioned in the letters and do not need to add hyperlinks to people, places, or events, or explain their historical background). Also, letters published on AiK do not have images, but they do have a "summary", which usually consists of either the first few lines of a letter or a short paragraph a blogger found interesting. Moreover, if available, they are enriched with information on sending and receiving location.

The complete processing of the transcribed letters has been recorded in handbooks by Laura Fahnenbruck, which are available for registered members on the AiK website. The handbooks allow all project members insight into the transcription, annotation, and publication guidelines and make them transparent at later stages of the project. Project members are also in close exchange with the project leaders Laura Fahnenbruck and Andrew Bergerson. At the beginning of their project membership and, if required, also in later stages, they receive an extensive introduction and training by Laura Fahnenbruck in all steps of the digitization and publication process.

For the success of crowdsourcing projects, the maintenance of close contact between coordinators and project members is one of the most important requirements (van Hyning 2019, 8). Using their experiences from the crowdsourced research project "Cinema in the GDR" as an example, Smolarski, Carius, and Plaul (2022, 311–12) point out the amount of extra effort and interaction it takes project collaborators to integrate the citizen scientists in a research project. A group of participants is heterogeneous and must be acquired suitably and should be motivated to stay for the entire duration of the project. Not only in citizen science projects but in general for crowdsourced projects, the quality of the work decreases when there is no long-term relationship between project leaders and the participants (Elmalech and Grosz 2017, 21). Moreover, interaction between researchers and citizen scientists is always bidirectional. Not only does it help to increase the quality of the work on the side of the citizen scientists, but it also provides valuable input, insights, and new perspectives for the researchers, as they might be confronted with challenging inquiries, reflections, and feedback from the project members, which could expand their understanding of the subject they are investigating (Smolarski, Carius, and Plaul 2022, 310). To do so, in their project Smolarski, Carius, and Plaul (2022, 314) expanded the opportunities for the citizen scientists to provide feedback from direct, verbal communication to quantitative and qualitative surveys, discussion rounds, and feedback letters, which allowed them to measure the quality and success of the participatory approach for both citizen scientists and researchers. T&S/AiK members are constantly encouraged to provide feedback on all procedures in the project.

Some of the researchers involved in the former T&S project and related projects within T&S's scope report how the collaboration with the interdisciplinary team of citizen scientists caused them to "reflect on [their] own iterative practices" (Baker et al. 2020, 63). It offered an opportunity for self-reflection on their own role as the people writing history and demonstrated how the correspondence served more as a way to reflect on one's writing about the past rather than as an actual historical source (Baker et al. 2020, 57, 68).

3.6 Overview of the Corpus

This chapter serves as an exploratory data analysis of the entire corpus. To our current knowledge, the correspondence encompasses a total of 2,626 letters by Hilde and Roland plus some more written by family or friends. It can be reasonably assumed that Hilde and Roland's family may continue to uncover missing pages. An example of this is a discovery of several missing pages from January/February 1944 in June 2023. 1,128 of the currently available letters were written by Hilde and 1,498 by Roland. In terms of tokens, the correspondence (excluding the rare letters from family members that were kept in the same folders) consists of 3,098,969 words. 1,539,489 of these were written by the Hilde, and 1,559,480 by Roland. Figure 3.3 shows the average letter length. The lower number of Hilde's letters can be explained by both gaps in the correspondence and, as the very similar total word count indicates, Hilde writing slightly fewer, but longer letters.

On average over all years and for both Hilde and Roland, a letter is 1,223 words long. While the initial letters exchanged between both spouses are relatively concise as they became acquainted, there is a gradual trend of increasing letter length until the year 1943. During his war captivity in Kiel, Roland's letters became gradually shorter. Conversely, during the last years of the war and days or entire weeks without receiving any letters, Hilde's letters became even longer. Steele (2011, 373) points out the increasing importance of writing and receiving

Figure 3.3: Average Letter Length per Year

letters for women at home as more time passed in the war and the situation both at home and at the front became worse. An increased need to talk to Roland might be one explanation for the steadily rising curve of Hilde's average letter length.

Figure 3.4 illustrates the letters' lengths, showing that there are some extremely long letters with more than 2,741 words. In total, the box plot depicts 47 outlier letters, with only 3 of them written by Roland; the remaining 43 were written by Hilde. Depicting the overall distribution of the number of letters in the entire correspondence could be misleading due to potential gaps in the correspondence or irregular furloughs. However, doing so is less problematic when focusing on Hilde's outlier letters (Figure 3.5). The number of outliers is relatively small in relation to the entire corpus (47 out of 2,626). The risk that there is still a large number of outlier letters not (yet) included in the corpus that would influence a plot too much is relatively small.

It might be anticipated that a plot of Hilde's notably lengthy letters by year would exhibit a concentration of such outliers in the final years of the war which contributes to distortion in the average letter length during that period, as shown in Figure 3.3. However, upon examining the outlier plot, it becomes evident that there are not more outlier letters at the end of the wartime period and therefore Figure 3.5 does not confirm the hypothesis of disproportionately long letters only in the last years of the war.

In order to further introduce the data, word clouds can be employed to visually present the most frequently occurring words. While the word clouds here are mainly supposed to give an overview of the corpus, they also show the most frequent words that later will be the most prominent features for the classifiers. The word clouds

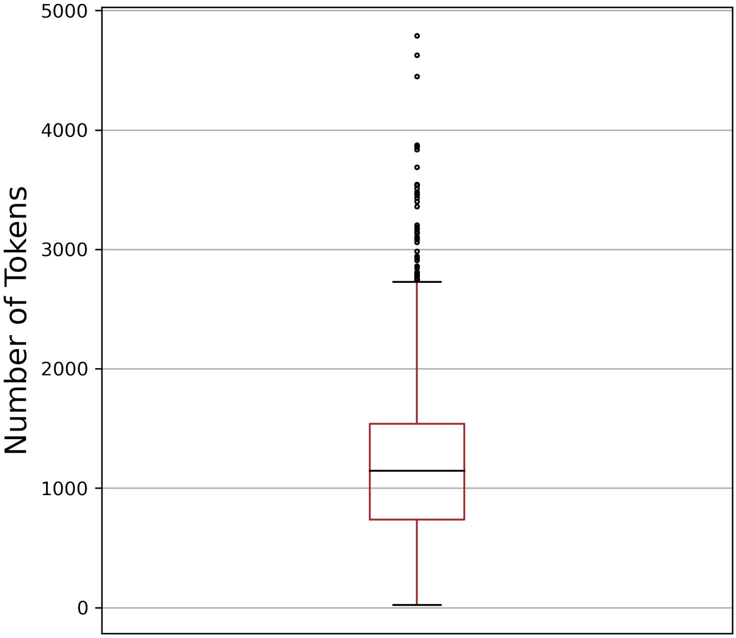


Figure 3.4: Letter Length (Box Plot)

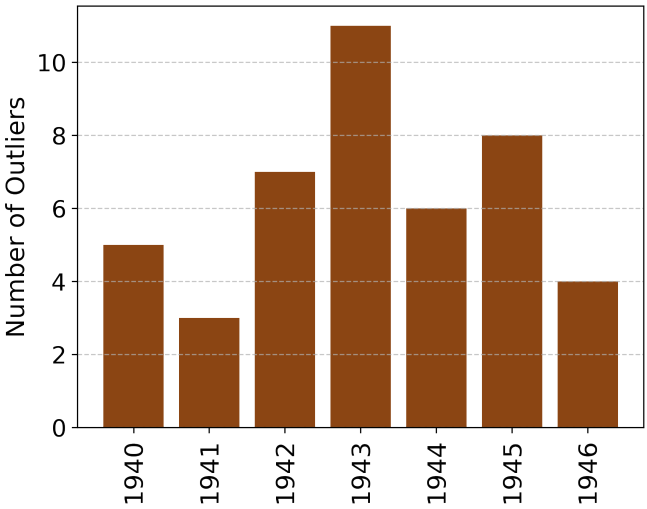


Figure 3.5: Outlier Letters by Hilde

(a) Word Cloud for Hilde

(b) Word Cloud for Roland

Figure 3.6: Word Clouds before extended Stop Word Removal

(a) Word Cloud for Hilde

(b) Word Cloud for Roland

Figure 3.7: Word Clouds after extended Stop Word Removal

in Figure 3.6 show the data after a basic stop word removal, including function words as well as weekdays and months (the stop word removal will be explained in Chapter 5.1 in detail).

The word clouds demonstrate how frequent the nicknames the spouses give each other are and indicate that they might have to be added to the stop word list. The updated word clouds with an extended, customized stop word list can be seen in Figure 3.7.

“Leben” is one of the biggest words, since along with the nicknames it is a word that Hilde and Roland use often to express their love for each other.

3.7 The Keywords

In addition to the textual content of the letters, which has just been presented, the keywords associated with the individual published letters play a supporting role in this master's thesis. Each published letter in AiK is labelled with several out of a total of 81 thematic keywords (how they were selected will be described below). The keywords ought to be perceived as concept terms that encompass a broader subject matter that Hilde and Roland address in their writing. The keywords' goal is to make the corpus better searchable and accessible for anyone interested. While a large part of the letters describes the spouses' everyday life or consists of expressions of mutual love, shorter and often less explicit parts focus on topics like ideology, war, criticism of the Nazi regime, or gender roles. The concept terms are designed to facilitate the process of locating specific segments and navigating through the corpus for both researchers and the public.

Their overall occurrence is plotted in the form of a word cloud on the website (Figure 3.8). The most frequent keywords on the 1,309 already labelled letters are "Gefühle" ("feelings"): 826 letters; "Familie" ("family"): 498 letters; "Schreiben" ("writing"): 476 letters; "Glaube" ("belief"): 428 letters; and "Feste" ("celebrations"): 328 letters.

When discussing keywords in the context of cultural and social studies, there is no getting around the concept of keywords introduced by Raymond Williams, established in the 1980ies. Intrigued by the polysemy of the word "culture" (as a formation of values vs. equivalent to society) (Williams 1985, 12–13), they started writing essays on a collection of words, which was later published as "Keywords: A Vocabulary of Culture and Society." A keyword, as defined by Williams, evolves with developing cultural and ideological meanings within a specific social discourse or context. Within their essays, they elaborate on the utilization of the keywords, addressing associated challenges and issues concerning their meaning (Williams 1985, 15). The keywords in Williams's collection originate from the broad field of culture and society and are supposed to be a domain-specific "vocabulary" rather than simply a "dictionary" (Durant 2008, 4), which would primarily be philological and etymological (Williams 1985, 17). In their critical reflection on Raymond's concept of keywords, Alan Durant discusses Raymond's selection of keywords, since the field of culture and society is broad and the selection has to be based on both linguistic and social criteria. According to Durant, Williams selected keywords talking "about the field of culture and society", which are polysemantic and commonly used in general and more technical contexts (Durant 2008, 5).

The relevance that Williams's keywords still or especially have today is mentioned by Heuser and Le-Khac (2012, 2): In the 1980s, Williams wanted to analyse an entire social discourse based on the keywords' development, but back then he simply lacked the methods to do so. Today, digital humanities scholars have the opportunity to trace these developments much more broadly and can explore long-term changes in language and discourse in large corpora.

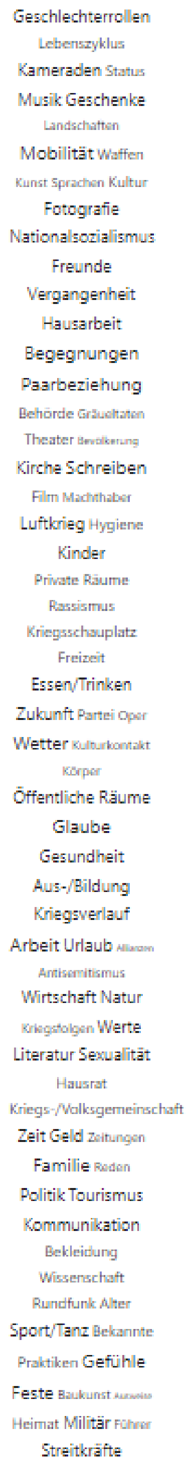


Figure 3.8: Word Cloud from AIK (Fahnenbruck et al. 2023)

Naturally, in the context of AiK the selection criteria are much narrower than those for keywords covering the entire field of culture and society. However, this has not always been like that. Before the changeover to AiK, in T&S there was no fixed set of keywords. The bloggers were provided with the straightforward directive to assign labels to a letter using as many words related to its content as possible. The sole stipulation was that these words should not appear verbatim in the letter, as such instances would already be discoverable through a full-text search. Like this, a keyword set of in total 1,689 words, some of them used very few times or even only once, was created (the entire list can be found in Appendix A.1). These many different and infrequent keywords hardly contributed to making the corpus better searchable. Similar to Williams's keyword selection process, which started from a large cluster of interrelated words and references (Williams 1985, 19), project leaders Andrew Bergerson and Laura Fahnenbruck combined as many keywords as possible to a smaller set of words describing overarching concepts when moving all letters from T&S's old WordPress blog to the new AiK website.

Before combining the old keywords as efficiently as possible into meaningful concept terms, they have not agreed on what these concept terms should be or how many of them are required to fully cover all topics in the correspondence. Rather, they emerged from the process of how it made sense to combine them. Durant (2008, 20) also states that there is no target number of keywords based on linguistic grounds but that it rather depends on "the interests, concerns or agenda of an anticipated readership." Ultimately, Bergerson and Fahnenbruck settled on a compilation of 81 keywords. The aim was not only to encompass as many words as feasible from the initial set of 1,689 keywords but also to provide readers of the blog an opportunity to search letters about any topic of their interest. Any letter that is blogged on AiK has to be labelled with one or usually more of the new keywords now.

Durant (2008, 17–19) defined the following criteria for likely keywords:

- They have to be used commonly and in day-to-day discourse.
- They are usually polysemous (different meanings in different circumstances).
- They are categorical in a way that they describe social/cultural concepts and practices.
- They are actively contesting in a social debate or dispute.
- They are part of a cluster, in which they commonly occur.

AiK's keywords meet most of these criteria as they are used in common parlance, attempt to describe subordinated concepts, and they all emerge from the narrow context of the letter correspondence. To address the polysemy, Laura Fahnenbruck wrote a glossary describing each keyword's particular meaning in the context of AiK, which is available on the website for all registered members. Moreover, on the website it is documented which original T&S keywords were included in which new AiK concept terms and which were not used at all anymore (see Appendices A.1 and A.2). Also, at least some keywords' contribution to a social debate or dispute can be assumed (especially those concerning societal and political issues beyond the writers' personal sphere).

Even though the final 81 keywords are now appropriate concept terms, the process of assigning them to the letters was problematic. Due to the large amount of time it would have taken to read all 1,309 labelled letters again, the new conceptual keywords were created to the best of the project leaders' knowledge. The previous class labels were consolidated into the reduced set without undergoing re-editing of the letters to which they were originally assigned. For example, all letters originally labelled with "Adresse" ("address") were reclassified under the new keyword "Schreiben" ("writing"). However, it remains uncertain whether "address" solely pertains to the location where the spouses sent their letters or if it encompasses other contexts, such as someone's new address after they moved or the address of a place they told each other about. Another example is the old keyword "Eid" ("oath") which was assigned to "Militär" ("military") but could emerge from a completely different context. No upper limit was set beforehand on how many old keywords should be combined into one new term. Therefore, some of the new terms are composed of a large number of old keywords. Consequently, a significant number of letters are labelled with these terms. Moreover, since in T&S, it was free to bloggers which keywords they chose to assign, it is not clear whether they chose them from Hilde and Roland's perspective (e.g., them explicitly making an oath), their perspective (the bloggers interpreting something as an oath), or future users' perspective (making a specific letter findable for users interested in the subject of oaths). Despite their imbalanced and inherently flawed nature, the creation of the set comprising 81 precisely defined keywords was essential to transform the keywords into classes that could be used for training a model.

In comparison to other editions of letters from the German-speaking world, it appears that the practice of rendering the individual documents searchable by keywords, as done in AiK, is relatively rare. The efforts to automate the keyword tagging through classifier training yielded unsatisfactory results, as evidenced by the performance of the applied classifiers discussed below. This has provided me with a deeper understanding of the challenges that might deter other editions from pursuing a similar approach. It can be assumed that other editions would face similar problems with noisy training labels since multi-class labelling (especially with such a high number of classes as 81) always gives room for variance and inevitably leads to noisy labels. In fact, editions that do use keywords comparable to those in AiK assign them manually.

The *Briefsammlung der Museumsstiftung Post und Telekommunikation* (Didczuneit and Goehle n.d.) uses a set of around 200 keywords,[5] which is more detailed than AiK's concept terms. Both place names and thematic keywords are assigned manually.

The digital edition of August Wilhelm Schlegel's correspondence (Strobel and Bamberg 2014) uses a smaller set of concept terms, similar to AiK. Their relevance in the corpus can be estimated via their size in a word cloud (Figure 3.9). The goal of these terms is to provide a "semantische Tiefenerschließung" ("semantic in-depth

[5] https://briefsammlung.de/feldpost-zweiter-weltkrieg/stichwort.html?keywordGroup=1

Figure 3.9: Word Cloud of *Digitale Edition der Korrespondenz August Wilhelm Schlegels* (Strobel and Bamberg 2014)

exploration") of the letters that meets the needs of an international scientific community. Apart from the keyword search, letters can be searched by epistolary metadata (sender, recipient, date), too. Together these strategies provide different layers of depth for a search (Strobel 2011, 149–50).

Vera Schwamborn worked on a digital edition of Vilém Flusser's letters after the philosopher's death in the early 1990ies. The digital edition is apparently no longer available online today, but their paper provides insights into the creation of an early digital edition aiming to make the letters searchable using keywords. The author was responsible for sorting the letters and tagging them with conceptual keywords. They report the scepticism of Flusser's widow regarding the breadth of the keywords. In the wife's opinion, the concept terms were too crude to cover her deceased husband's thinking (Schwamborn 2017, 1). This highlights one of the problems with conceptual keywords. In the AiK correspondence, especially Roland uses metaphors and figurative language. Concept terms always carry the risk of not being fully able to represent these appropriately.

Other letter editions from the German-speaking world, such as *Arthur Schnitzler Briefe* (Müller, Susen, and Untner 2018), *Auden Musulin Papers* (Mayer et al. 2023), and *Ernst Haeckel Online Briefedition* (Breidbach and Bach 2023) use entities mentioned in the letters (persons, works, institutions, places, events) or epistolary metadata (sender, recipient, date) to make the letters searchable, but do not employ thematic keywords.

3.8 Geolocating Places in the Corpus

Even though the focus of the thesis is on thematic keywords, for the sake of completeness AiK's GIS approach should be introduced briefly. It is another way to make the letters more accessible. The goal is to, if possible, geolocate a letter by adding the coordinates of its sending and receiving location, as well as a list of all the places Hilde and Roland mentioned in it. Currently, the AiK bloggers do this manually and no named entity tagger is used for obtaining place names. When comparing AiK with other letter editions, it seems that using named entity recognition (NER) is still not a common practice when editing correspondences. An edition similar to AiK is the *Briefsammlung der Musumsstiftung Post und Telekommunikation* (letter collection of Museumsstiftung Post und Telekommunikation) (Didczuneit and Goehle n.d.). Unlike AiK, they do not publish complete correspondences, but rather selected letters written by various authors and from several contexts (field post from World War II or letters from 18th and 19th century, World War I, and GDR). Moreover, they only publish letters written by soldiers on the front line and no letters written from home. However, similar to AiK, they add the sending location and thereby make letters searchable by places. For instance, all letters written from places in the German Reich can be filtered by city/town (Figure 3.10).

Given that the website lacks information regarding the methodology behind the acquisition of these sending locations from the letters, I reached out to the editors, Dr. Veit Didczuneit and Gunnar Goehle, via e-mail to inquire whether they employed a NER-tagger. Their response was the negative. Similar to AiK, the process of manually adding locations to a list of places mentioned in the letters is performed by the transcribers. Also, for the letters' transcription, no transcription software is used but a team of researchers manually transcribes all letters and mutually proofreads the transcripts.

One advantage of geolocating the places in AiK is that they can also be displayed on a map (Figure 3.11). Following the annotation of a letter by bloggers with mentioned places, either the bloggers themselves or another project member disambiguate these locations by assigning coordinates. The coordinates then are employed to geolocate the places on the map, created with the JavaScript library Leaflet.[6] This feature enhances the searchability of the letters. Users can locate them not only by searching for a specific place in the search bar but also by accessing them through the map. The geolocating is still in progress. The following section of the map shows its status as of August 2023. There are, of course, a large number of other letters that are just not yet noted on this version of the map.

In many letters, the sending and receiving location are not explicitly mentioned. Still, it is often evident from the context that Hilde sent her letters from her hometown of Oberfrohna. Sometimes she visited Roland's parents in their home town of Kamenz, but in this case, both of them would mention it explicitly in their letters. Roland was usually at his current deployment, like Thessaloniki, Bucharest, Constanța, or Kiel. When travelling, Hilde usually mentioned her destinations.

[6] https://leafletjs.com

Figure 3.10: Places in *Briefsammlung der Musumsstiftung Post und Telekommunikation* (Didczuneit and Goehle n.d.)

For Roland, on the other hand, it was often not possible to explicitly mention places due to censorship, which made him use initials or abbreviations instead. Prohibiting specifying places in the letters should prevent enemy powers from concluding about, for instance, soldiers' deployments (Humburg 2011, 81).

Figure 3.11: Map of *Alltag im Krieg* (Fahnenbruck et al. 2023)

Chapter 4

Methods and State of Research

4.1 Objective

Located in the field of digital humanities, the goal of this master's thesis is to merge a traditional humanities research question with a machine learning problem. Given that the data under consideration are textual, this study falls within the domain of Digital History and Distant Reading. Distant Reading is a term that was established by Franco Moretti in 2000. It describes approaches to analysing large corpora of text that go beyond the close analysis of individual subtexts within the corpus. In contrast to Close Reading, Distant Reading approaches texts in a manner in which information is conveyed not solely through the plain input text, but also by form, genre, and themes (Moretti 2013, 48-49). In this thesis, analysing the keywords in the corpus of letters which are partially obtained by a text classifier will function as a Distant Reading method. Creating line charts that depict the frequency of specific topics in Hilde and Roland's correspondence allows to establish a connection between the keywords and real-world events during World War II. This effort aims to place the individual letters in a broader temporal context (political and societal changes in the German "Reich" during World War II) as well as spatial context (by highlighting the dichotomy between Germany and the eastern front, where the spouses were situated).

By applying text classification and interpreting the obtained class labels quantitatively and qualitatively, the thesis showcases the potential of the combination of traditional and digital humanities to approach questions in digital history and the field of digital editions. While quantitative methods in the humanities as a consequence entail changes in qualitative research (Cohen and Rosenzweig 2005, 84), it remains uncertain whether these methods can consistently capture what Heuser and Le-Khac (2012, 2) call the "nuance and complexity we value in the humanities."

Hence, it must be emphasized at this point that the classification of the letters and an examination of the resultant class labels alone cannot constitute an interpretation of the correspondence. Interpretation becomes possible only through contextualization and a highly critical reflection of the keywords. Digital methods in the humanities such as Distant Reading should always only be seen as an additional method to, for instance, Close Reading, but never as a replacement. The results of both kinds of studies are justified, both require thoughtful scrutiny of the researcher, and one cannot be preferred over the other. Instead, Distant Reading provides an additional methodology to approach research questions (Jockers 2013, 29–30).

Scherstjanoi (2011, 122), who argues for a very cautious use of field post letters as a historical source, does concede that mass analyses of them can be useful for questions of a general nature, as long as their focus is not too narrow.

Digital history allows researchers to work on big amounts of data to support hypotheses and find correlations among sources of the past (Au Yeung and Jatowt 2011, 1231). Due to the size of the Oberfrohna corpus of around 3 million tokens, digital methods facilitate the work with the entire corpus. Furthermore, studying the temporal fluctuations in topics of interest through Close Reading would demand a significantly greater investment of time and labour compared to the efficiency achieved by utilizing the automatically assigned keywords. However, it must be emphasized that the utilization of quantitative methods brings a shift in focus to some extent. Purely qualitative research questions about the temporal variance of topics in the correspondence might not draw on the keywords and would be formulated differently, aiming at different goals.

One of the main assumptions of Distant Reading is that contextualizing individual texts within a larger framework enhances the interpretability of the individual texts themselves (Jockers 2013, 25). By placing individual letters of the Oberfrohna correspondence in a greater context using the keywords, one can ask specific questions within this contextual backdrop: What does the rising or falling frequency of a certain keyword reveal about the topic's relevance for Hilde and Roland, or about how Hilde and Roland perceived it over time? According to Jockers (2013, 27), Distant Reading allows "contextualization on an unprecedented scale."

Since the analysis of the variance of topics over the course of the war represents a link between qualitative and quantitative methods, the question of temporality in the use of Distant Reading and digital history arises. In their article "Die Modellierung des zeitlichen Vergleichs als Kernkompetenz von Digital History?" ("The modelling of temporal comparison as a core competence of Digital History?"), Katrin Moeller (2022) raises the question of how to deal with discontinuities in timelines when applying quantitative methods.

Gaps in the timeline and infrequencies are in fact a problem in the Oberfrohna correspondence. Figure 4.1 shows the available letters per month and, therefore, how the amount of information we have about each month in the correspondence varies. (It has to be kept in mind that the Figure 4.1 only illustrates how many letters the corpus contains, but not how many letters Hilde and Roland actually wrote, as some of them might have gotten lost). When analysing time series, the first question to ask is whether they are even comparable (Moeller 2022, 96). Figure 4.1 demonstrates the challenge inherent in conducting a comparative analysis of keywords from letters authored before Roland's deployment with those from the subsequent months and years of separation of the spouses (starting in July 1940). For most of the time after July 1940, Hilde and Roland's communication took place in the letters, therefore a much better overview of the topics they were concerned about is possible. Even when comparing the frequencies of keywords over time relative to the number of letters, due to the shift from partially verbal and partially written to almost exclusively written and preserved communication, this comparison can never be unbiased. Moreover, there are gaps in the corpus, for instance for Hilde in January/February 1944 and May/June 1945, when Roland did not receive her letters. Moeller (2022, 91) advocates for innovative solutions

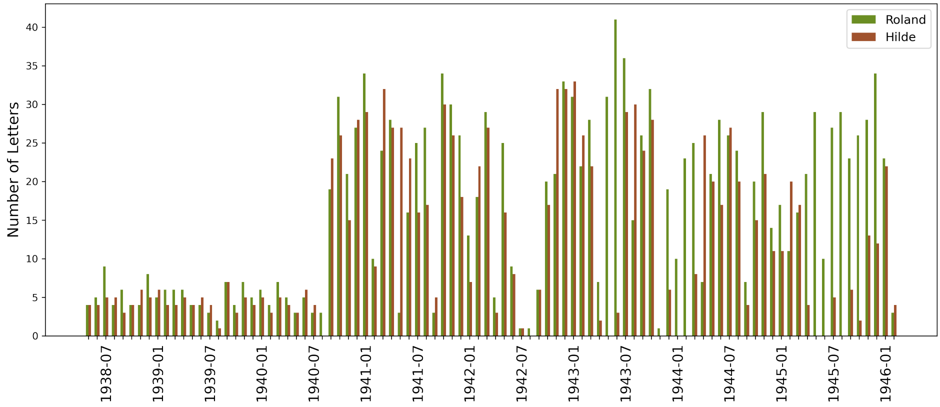


Figure 4.1: Available Letters per Month

to address source losses in digital history, mirroring the approaches embraced in classical history: One way to deal with source losses they suggest is to fill them with estimations and extrapolations. Another way to handle temporal gaps is to explicitly single them out because they are just as much a part of the corpus as the available documents are. Considering the imprecision of the keywords that can be obtained by the classification model, the idea is not to add even more uncertainty to the analysis by estimation but to treat missing letters as legitimate parts of the historical source.

4.2 On the "Digital" Side: Text Classification

4.2.1 Introduction to Text Classification

The first objective of the thesis is to predict the keywords of the letters as accurately as possible. To do so, a Logistic Regression model, an SVM, a bidirectional LSTM with CNN, a rule-based model, and a combination of some of these models with majority voting will be applied. Before these models are introduced in particular, this chapter provides a general introduction to text classification.

Text classification can be defined as determining to which class (in case of binary or multi-class classification problems) or classes (for multi-label classification problems) a document belongs. In machine learning, this makes it a supervised task, since a supervisor has to define classes and label the documents accordingly to direct the training process (Manning, Raghavan, and Schütze 2009, 253). For semi- or distantly supervised classification tasks, the classes and/or labels are created automatically. For instance, Phan, Nguyen, and Horiguchi (2008) use topics obtained by applying topic modelling as classes to train the classifier. Keith et al. (2017) automatically extract noisy class labels from a knowledge base and attempt to improve their quality. The classification task for the Oberfrohna correspondence is strongly supervised since labels from human annotators are treated as gold labels.

The output of a classification task is either a discrete label or a numerical score (the probability that an instance belongs to a class) that can be converted to a discrete label using a certain threshold (Aggarwal 2015a, 2). It must be noted that the existing class labels for the data have to be treated as gold labels for practical reasons, even though a qualitative evaluation promptly reveals that they might not consistently hold up to this standard. Due to the error-prone training and evaluation labels, the models can only achieve a certain upper limit of performance and a perfect classification of the test data will not be possible—or at least not measurable.

Classifiers can be applied on different levels of a text: on sub-sentence level (portions of sentences, e.g., parts of speech), sentence level (portions of a paragraph), paragraph level (portions of a document) and document level (seeking all categories within a complete document, such as the letters in this study) (Kowsari et al. 2019, 2). While at first text classification has mainly been applied in information retrieval to make large collections of documents searchable, nowadays it is used in many different domains (Kowsari et al. 2019, 55). Examples from the literature include information filtering (selecting or rejecting relevant or irrelevant information within the data), recommender systems based on user feedback or search queries (Kowsari et al. 2019, 55–56), spam filtering, filtering of, for instance, sexually explicit content, sentiment analysis (classify reviews as positive or negative), and topic-specific search (Manning, Raghavan, and Schütze 2009, 253–54). Topic-specific search seeks for more general, larger topics in a document (Manning, Raghavan, and Schütze 2009, 254), just as the topics represented by the keywords in AiK.

Kowsari et al. (2019, 2) define the following crucial steps in a text classification task:

(a) feature extraction
(b) dimensionality reduction (for instance principal component analysis, not applied for this thesis)
(c) selection of an appropriate classification model
(d) the evaluation of the model

For training, the data are represented by a feature set defined in the steps of feature extraction and dimension extraction. Feature matrices, capturing the features of each document, are commonly referred to as X_{train} or X_{dev}. The labels are stored in label matrices y_{train} or y_{dev}.

During the training phase, the classifier learns from the extracted features representing the individual documents and their associated labels. In the prediction phase, the classifier is supposed to predict the label of the development data, using the same feature set as in the training phase (Tang, Alelyani, and Liu 2015, 40).

Of the four steps they defined, for Kowsari et al. (2019, 1–2) the selection of the model is the most important one, but also a big challenge for researchers. The selection of the right model depends on the trade-off between its performance and its complexity, as well as on the application (Gunasekara and Nejadgholi

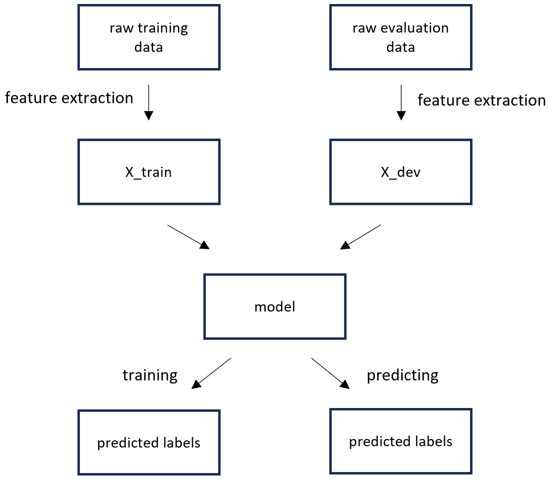


Figure 4.2: Basic Architecture of a Classification Model

2018, 21–22). A successful algorithm is supposed to understand complex data with non-linear relationships of their features (Kowsari et al. 2019, 1). In the case of the Oberfrohna correspondence, the model must be capable of dealing with the multi-label and multi-class classification problem as well as the noise (see Chapter 4.2.2). Compared to binary classification, multi-label classification is a much more challenging problem (Gunasekara and Nejadgholi 2018, 21). The approach to face this challenge in the experiments for this thesis is to treat the multi-class/multi-label classification problem with 81 different classes as 81 binary classification problems, separating letters labelled with one particular class from all other documents which are not labelled with this class. According to Tsoumakas, Katakis, and Vlahavas (2006, 3–4), learning L binary classifiers (one for each label l in the label set of length L) is the most common transformation method for multi-class problems. The dataset is split into L different datasets, whose documents are labelled as positive or negative for a specific class. In Scikit-Learn, the MultiOutputClassifier module allows to fit one classifier for each label when using an algorithm implemented in the framework[1] (such as Logistic Regression or SVM, which will be used for labelling the letters of the Oberfrohna correspondence) (Pedregosa et al. 2023).

Both rule-based systems and models based on machine learning can be used for text classification. Rule-based models use rules created by human experts. Often, these models' advantage is high precision, but creating the rules might involve a lot of manual labour and require high linguistic expertise. Algorithms using machine learning, on the other hand, can learn independently from the training data and can also cover cases to which no rule applies (Roth and Rocha Souza, 2021). For this thesis, both a simple rule-based classifier as well as machine learning models are used.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html

Machine learning models in text classification can be categorized broadly into four different types: linear models, support vector machines, decision trees, and neural networks (Tang, Alelyani, and Liu 2015, 40). The models selected will be introduced below. A good general entry point of these models' application for multi-label tasks is provided by e.g., Tsoumakas, Katakis, and Vlahavas (2006).

When different models are used to classify the data, they achieve different results and vary in their overall performance, precision, recall, bias, and capability to deal with noise. To combine their strengths and avoid their weaknesses, ensemble techniques such as stacking can be employed to improve overall performance (Ikonomakis, Kotsiantis, and Tampakas 2005, 972; Gunasekara and Nejadgholi 2018, 23–24). For some of the most popular approaches for stacked classifiers, Bi et al. (2004, 127–28) refer to Sebastiani (2001), who describes *majority voting* (the majority of all combined classifiers predicting a class), *weighted linear combination* (the weights of all classifiers are combined), *dynamic classifier selection* (dynamically selecting the classifier with the highest local accuracy), and *adapter classifier combination* (summing all classifiers together). For this thesis, the simplest approach, majority voting, will be applied.

4.2.2 On the Impact of Noise in the Data

Several sources of noise might negatively affect the performance of the classifiers. Sources for noise in the corpus are both the text itself and the letters' class labels. Some of the noise in the text data is caused by the automatic transcription with Transkribus. However, this affects only the letters which have not been proofread by the time of their export from Transkribus and the old T&S blog in December 2022 (around 35% of the entire corpus; see Chapter 5.1). The letters used for training and evaluating the classifier have been proofread, therefore HTR errors will not affect the training but only the labelling of the unknown letters for interpretation. Noise on the labels is mainly caused by how the new concept term keywords (which serve as class labels now) were created based on the old keywords (as explained in Chapter 3.7). It makes the resulting class labels imbalanced and leads to classes that are ambiguous and hard to distinguish for a text classifier. Moreover, even the quality of the original keywords before they were combined into concept terms is questionable and introduces noise in the data.

HTR transcription errors The first factor that might influence any model's performance on the unknown letters is the noise in the text data. By the time of the export of the corpus in December 2022, the letters until August 1943 had been proofread at least once (around 65% of all letters). The remaining letters still contain HTR transcription errors. Since OCR/HTR systems tend to produce similar substitution errors over the entire dataset, some of these can be corrected systematically (Agarwal et al. 2007, 5). I manually replaced recurring errors but there are still many more in the not proofread transcripts.

Surprisingly, the studies of both Agarwal et al. (2007) and Apostolova and Kreek (2018) indicate that noisy text is less of a problem for classification. Agarwal et al. (2007, 11) used corrupted datasets to train a Naïve Bayes classifier and an SVM.

They tried different combinations of clean and noisy training and evaluation sets. The experiment with a clean training set and noisy evaluation set is what comes closest to the data of the Oberfrohna correspondence: the classifier is trained on the proofread data but should be applied to the raw transcripts. For clean training data and noisy test data, they report only a slight drop in accuracy compared to clean training and clean test data, depending on the level of introduced noise. There are minimal differences in the accuracy using datasets corrupted with 0%, 40% or 70% noise. Only at 100% noise (a spelling mistake in 100% of all words), the accuracy drops significantly (Agarwal et al. 2007, 8). Also, Apostolova and Kreek (2018, 108) report almost no decrease in performance up to 50% artificially introduced noise. Even though the different algorithms they tried varied in their performance depending on the level of noise in the training data, for all of them the degradation of performance on a clean test set was small. However, they also report that when measuring the classifier’s performance on the noisy training dataset via cross-validation, the results are “typically over-pessimistic”. Unlike the results reported by Agarwal et al. (2007), which did not indicate a large decrease in performance on a noisy evaluation set, this does suggest that the performance of the classifier which was trained on the proofread letters will decrease on the raw transcripts with HTR errors.

A paper dealing with a similar problem was written by Mohammad (2018). To train a classifier on toxic comments, the author evaluates the advantages of several approaches to cleaning data from noise. These comments usually contain many spelling mistakes, on the one hand caused by typos and, on the other hand, to dodge automatic toxicity filters (Mohammad 2018, 1)—just as the letters contain HTR transcription errors. Non-standard language is a less relevant problem for training a classifier on the Oberfrohna correspondence in comparison to the study from Mohammad (2018) since it can be trained specifically on Hilde and Roland’s dialectal and idiolectal terms. The author applies several transformations to the data and trains four different models (Logistic Regression, Naive Bayes with SVM, Extreme Gradient Boosting, and a Bidirectional LSTM with FastText embeddings) in order to compare performance gains (Mohammad 2018, 4). Improved model performances are expected since the transformations should result in feature reduction. Leaving all tokens just as they are would let the model treat tokens that stem from the same word as different features, resulting in feature explosion and increasing difficulty in training the model. The transformations done by the author include the removal of rare words; writing out blacklisted words (e.g., “sh**”, in which some characters have been replaced by asterisks) with regular expressions (comparable to the systematic correction of frequent transcription errors); matching infrequent words with proper names, such as cities and persons; and replacing misspellt words with the correct spelling by fuzzy matching based on the Levenshtein distance (Mohammad 2018, 2–3).

For English texts, the WordNet algorithm allows an even more enhanced feature reduction by not only normalizing spelling but also organizing words in sets of synonyms (Miller 1995). As an enhancement to lemmatizing/stemming, WordNet would enable a semantical rather than morphological mapping of words to features. However, its main issue is the polysemy of many words making. This makes it

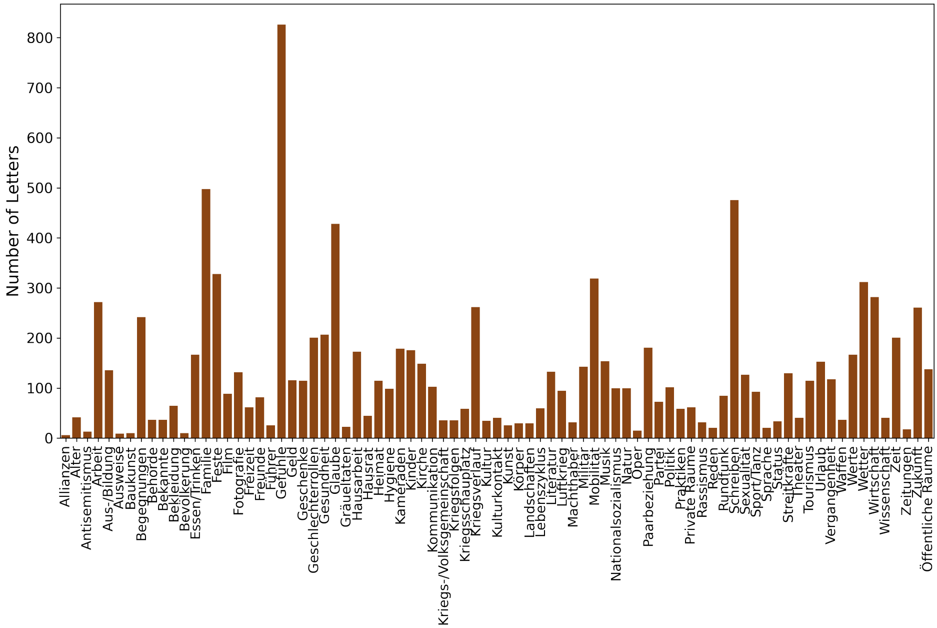


Figure 4.3: Distribution of Classes on Labelled Data

difficult for the algorithm to disambiguate them correctly and is likely to result in noisy features (Scott and Matwin 1999, 6). A German alternative for a semantical wordnet is GermaNet (Hamp and Feldweg 1997; Henrich and Hinrichs 2010). The approach will not be pursued further in this master's thesis to avoid the risk of introducing additional noisy features. Furthermore, the corpus contains many unknown (dialectal and idiolectal) words that might not have been captured in a wordnet. Additionally, the results from (Mohammad's 2018, 5) study indicate that extensive data cleaning was not as useful as expected by the author. Especially for Logistic Regression and the combined Naive Bayes-SVM, most transformations caused a decrease in accuracy. However, for the LSTM network with FastText, the fuzzy matching of words to correct spelling mistakes resulted in an increased performance.

Class imbalances The distribution of keywords over the labelled letters is imbalanced, with some keywords applied to as many as 825 out of 1,309 letters, while others are assigned to as few as 5 letters. Scikit-Learn requires that each class has at least one instance in the training, development, and test set (Pedregosa et al. 2023). Due to the chronological split of the labelled data into these three sets (see Chapter 5.1), certain class labels with a very low frequency were not present in every split of the complete labelled dataset. To address this, I manually selected and annotated 5 unlabelled letters and added them to these sets. Figure 4.3 shows the unequal distribution of the classes across all labelled letters. This class imbalance arises from both the selection of the particular thematic keywords that should make the corpus searchable and the problematic combination of the old keywords. Some topics were more relevant for Hilde and Roland than others, leading them to write more about these subjects. Other topics might have been relevant, too, however, Hilde and Roland avoided mentioning them in their letters due to internal or external censorship. All these aspects result in imbalanced classes. Moreover, some uncertainty remains regarding whether a substantial proportion of the annotations is indeed correct. As a consequence, it is unclear if the distribution depicted in Figure 4.3 truly serves as an accurate representation of the topics addressed within the letters. As the keywords are now clearly defined on the AiK website, it can be assumed that the new labels that bloggers will assign to the currently unlabelled letters will be more consistent (as far as this is possible, considering the annotator disagreement in such a label set, which will be addressed below) and would provide better training, development, and test data for a model.

Semantic and thematic overlaps of keywords Another issue about the labels that has to be highlighted are potential semantic overlaps of keywords. Inspecting these overlaps suggests that maintaining clear distinctions between classes is challenging. For examining multi-class classifiers trained on noisy data, Agarwal et al. (2007, 7, 11) chose the Reuters-21578 dataset (consisting of Reuters news articles) with 90 different classes. Along with labelling challenges similar to those T&S/AiK has to face, the authors observed a separability problem arising from the large number of classes. The overlap between classes turned out to degrade the overall results on the SVM classifier they applied to the data. For AiK, one potential approach to address this issue could involve enhancing the differentiation of vague or ambiguous classes, for instance “Kommunikation” (”communication“), which is defined as the following: “Umfasst zwischenmenschliche Kommunikation in Gesprächen und am Telefon, sowie Kommunikationsmodi wie Geheimnisse, Klatsch, Streitigkeiten, Koseworte, Witze, Gerüchte, Lästern, nicht aber die schriftliche Kommunikation und postalischer Verkehr (siehe Schreiben).” (“Includes communication through spoken conversations, including telephone conversations, as well as modes of communication such as secrets, gossip, arguments, term of endearment, jokes, rumours, blasphemy, but not written communication and postal correspondence (see Writing).”) (Fahnenbruck et al. 2023) While per definition the topic of writing is differentiated from oral communication, in practice nearly half of the letters in the labelled portion of the corpus according to their keywords cover not only oral but also written communication (see Figure 4.4).

The descriptions of the keywords “Oper” (“opera”), “Musik” (“music”), and “Theater” (“theatre”) (Fahnenbruck et al. 2023) also suggest thematic overlaps:

- Oper: “Umfasst auch Operette, Opernsänger und Opernsängerinnen, den Besuch einer Oper.” (“Also includes operetta, opera singers, attending an opera.”)
- Theater: „Umfasst auch Theaterbesuche, Theaterschauspieler und -schauspielerinnen, Fronttheater, nicht aber Oper und Film.” (“Also includes

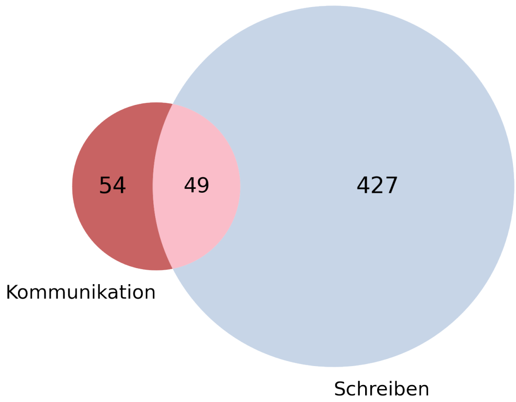


Figure 4.4: Overlap of “Kommunikation” and “Schreiben”

theater attendance, theater actors and actresses, front theater, but not opera and film.”)

- Musik: “Umfasst auch Musiker, Noten, Dirigenten, Konzerte, Sängerinnen, nicht aber Singestunde und Kirchenchor (siehe Kirche), Wunschkonzert (siehe Rundfunk) und Oper (siehe Oper).” (“Also includes musicians, sheet music, conductors, concerts, singers, but not singing lesson and church choir (see church), Wunschkonzert (see radio) and opera (see opera).”)

Still, relatively few letters share all labels, as Figure 4.5 shows.

In the AiK project, there are currently no resources for re-labelling already published letters and differentiating ambiguous categories, as the focus is on proofreading and blogging of the so far un-edited letters. In this thesis, the presence of noise in all training, development, and test data must be accepted as a given. The continuous inconsistency across all segments of the dataset establishes a form of consistency in itself.

Lacking quality of the original annotations The original keywords assigned by human annotators are not free of noise either. Suissa, Elmalech, and Zhitomirsky-Geffet (2021, 271) distinguish two ways of dataset generation by human annotators: crowd-based (for easier labelling tasks that fall under common knowledge) and domain expert-based. The peculiarity of the labelling task in AiK stems from the fusion of these two aspects: while anyone can contribute to the project, possessing domain knowledge (background understanding of Hilde and Roland’s story, along with comprehension of each conceptual keyword within the context of the AiK project) is essential for accurate labelling. Comparing different strategies and reward schemes to improve the outcomes of crowdsourcing, Huang and Fu (2013, 621) identify quality control as “one of the biggest issues”. There is no “real” ground-truth of the AiK labels that human annotators assigned and their

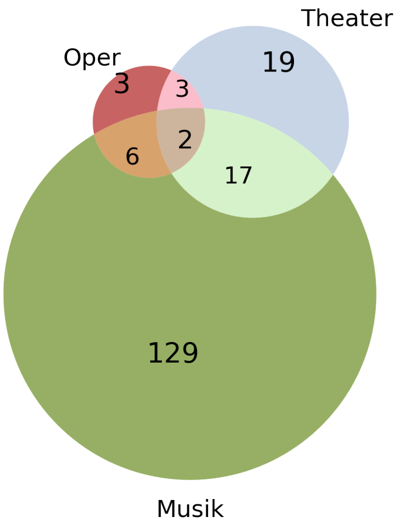


Figure 4.5: Overlap of “Oper”, “Musik”, and “Theater”

quality cannot be assessed, but it can be assumed that the quality of labels improves as project members process more letters and gain a deeper understanding of the domain. Huang and Fu (2013, 621, 628) figure that for the quality of crowdsourced annotations, social transparency plays a significant role, such as, for instance, sharing demographic information or interacting with teammates. The constant interchange with the project leaders and the other project members is an indispensable part of the AiK project and therefore inherently indicates a high quality of the work. The bloggers who annotate the letters usually do not get paid. Studies (Elmalech and Grosz 2017, 22; Huang and Fu 2013, 622) even suggest that financial benefits do not contribute to an increased quality of work and primarily influence the number of tasks accomplished by a worker. Quantity, however, is a focus that does not align with the decelerated way T&S/AiK aim to write history.

It is important to bear in mind that all labels assigned by human annotators to the Oberfrohna letters are based on a subjective assessment. Another person re-viewing them could disagree or even add more labels. Schwamborn (2017, 3), who manually labelled letters of the Flusser Edition with concept terms, argues that this subjectivity makes the keywords vivid: it places people in dialogue with the texts and brings in different associations and viewpoints that vary daily. When labelling complex texts, fuzziness is always present and adds a human factor to the plain annotation task. Although this approach is logical within the context of the humanities-oriented labelling task, it poses challenges when the labels are intended for training a classifier and must be regarded as ground-truth for this.

Arhin et al. (2021, 3) examined three annotated ground-truth datasets of toxic text data in terms of their inter-annotator agreement (consistency between multiple persons labelling the same document) and intra-annotator agreement (consistency between one person labelling one document multiple times). Inconsistencies in labelling often stem from missing context and ambiguous annotation guidelines. These inconsistencies have repercussions on the model's performance, highlighting the importance of clear instructions for annotators (Arhin et al. 2021, 2–3). Given that the context of the letters is well-defined for the project members, missing context is not a significant concern in the AiK project. Clear labelling instructions, on the other hand, which were missing completely in T&S, are now provided through the detailed definitions of the 81 keywords in AiK. However, these do not affect the already existing labels which were created previously and now serve as ground-truth for the classification model. One notable advantage of the annotation with multiple labels rather than binary labels is that it can result in less noisy datasets and improve classifier predictions (Arhin et al. 2021, 9).

Due to the lack of clear keyword definitions in T&S, the relatively extensive label set of 81 keywords, which inherently allows for the possibility of semantic overlaps between keywords, and the existence of ambiguous terms, it is reasonable to anticipate instances of intra- and inter-annotator disagreement. Arhin et al. (2021, 9) state how inter-annotator disagreement increases with an increasing number of annotators. However, studying the impact of noise in data and labels on classifiers, Agarwal et al. (2007, 11) carried out a statistical ANOVA (Analysis of Variance) and Gage R&R (Repeatability and Reproducibility) test that revealed that 53% of the time, multi-labelling results could not be reproduced amongst various annotators and 35% of the time not by the same annotator. For a benchmark of inter-annotator consistency, Agarwal et al. (2007, 11) refer to a presentation from Lewis et al. (2003), according to which 30% of inter-annotator disagreement can be accepted. For Arhin et al. (2021, 10), intra-annotator consistency is a sign of high dataset quality, while they interpret inter-annotator disagreement as an indication of annotator diversity, ensuring the representation of minority groups in prediction tasks.

4.2.3 The Models

As previously mentioned, the models that will be used in the experiments of this thesis are a rule-based model, Logistic Regression, SVM, and a bidirectional LSTM with CNN layer, as well as a combination of the three best models with majority voting to decide on the correct class. Especially Logistic Regression and SVM are standard and easy-to-use linear classification algorithms. Linear models in general assume that the instances of two classes are linearly separable. At this basic conceptual level, all different linear classifiers are very similar to each other, even though their implementation is different (Aggarwal and Zhai 2015, 315). If there is indeed a linear relationship between the instances and their features, linear classification models will be able to find a perfect classification. However, usually instances with different labels can share many of the same features. As a result, the training data are not linearly separable. Incorrectly labelled instances are punished with a loss. The overall goal of the classifier is to minimize the loss. Along with the neural network, Logistic Regression and SVM will be introduced in the following:[2]

Logistic Regression Logistic Regression is a more sophisticated version of the classical Perceptron algorithm. The Perceptron is a traditional linear, discriminative model. During the training process, this algorithm learns a weight for each feature and iteratively improves the feature weights by adding + or -1. The class of a document is predicted by a score based on the sum of all feature weights multiplied by the feature value (for instance negative class if score < 0, positive class if score > 0). The loss is determined solely by comparing the signs of the scores. If the signs of the score and the actual label differ for an instance, the loss for that instance is 1. If they match, the loss is 0.

Instead of producing only signs, the output of Logistic Regression is the probability that an instance pertains to a specific class. While multinomial Logistic Regression can have three or more possible outcomes, binary Logistic Regression can classify instances into either a positive or a negative class. For assigning the keywords to the Oberfrohna correspondence, for each of the 81 classes, a binary Logistic Regression model is trained, determining whether an instance belongs to a certain class or not. The output of binary Logistic Regression falls between 0 and 1 and a threshold (e.g., 0.5) must be defined to determine whether an instance belongs to a class (Aggarwal and Zhai 2015, 312).

Just as for the Perceptron, for Logistic Regression in the first step numerical scores (logits of the probabilities) are calculated by multiplying feature values with weights that are iteratively improved (Linear Regression) and may involve the addition of a bias term to increase or decrease the output value. To iteratively find the best weights, an optimization algorithm is employed (e.g., stochastic gradient descent, see Jurafsky and Martin (2023, 11)). In Scikit-Learn, the optimization can be defined via the solver parameter. Regularization prevents large feature weights. Large weights would cause overfitting on the training data, leading to poorer performance on test data. Logistic Regression provides two regularization options: L1 (Lasso) and L2 (Ridge) regularization. L1 regularization adds the absolute values of the weights to the loss function as penalty terms. The weight matrix becomes sparse as smaller and less important weights are set to zero. Therefore, L1 serves as a feature selection technique for datasets with a huge number of features. Genkin, Lewis, and Madigan (2007, 302) point out that thereby Lasso regularization enables state-of-the-art text categorization effectiveness and efficient, spare models. With L2 regularization, on the other hand, squared weights are added to the loss function. The resulting weight matrix is not sparse and all features are kept (Nagpal 2017).

In the second step, the scores are rescaled and normalized to probabilities. To transform them into probabilities, they have to be passed through the sigmoid function:

[2] Unless otherwise cited, the information about the widely used machine learning techniques and neural networks in this section is based on my own knowledge acquired during the course “VU Practical Machine Learning for Natural Language Processing”, offered in the summer term of 2021 by Benjamin Roth and Renato Rocha Souza at University of Vienna.

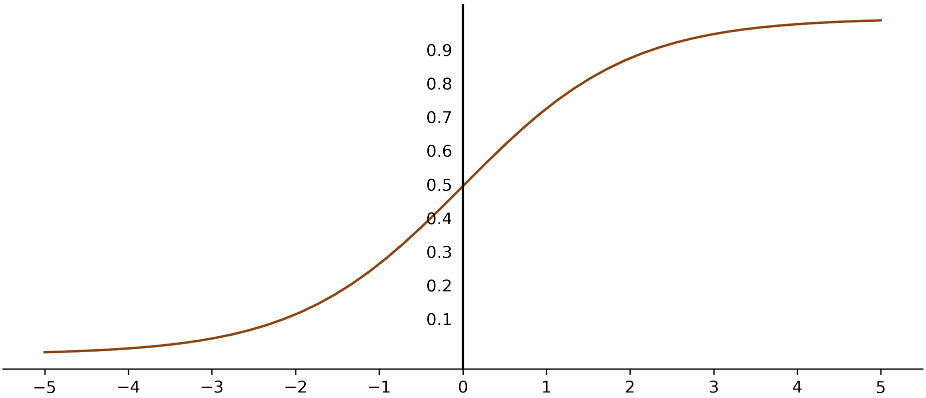


Figure 4.6: Graph of the Sigmoid Function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

where:

$\sigma(x)$ = probability of an instance belonging to a class

x = linear output score

The sigmoid function makes the scores for all classes positive and the sum of all probabilities over all classes (in the case of the binary Logistic Regression problem in this thesis for the positive and the negative class) becomes 1. The sigmoid/logistic curve defined by the sigmoid function is s-shaped (Figure 4.6). For determining the weights, Logistic Regression uses a logistic loss, defined by the probabilities of a class. Similar to the Perceptron, the loss is smaller if the sign is correct, but it can still decrease as predictions become closer to the correct class.

The utilized loss function is known as cross-entropy or negative log-likelihood. Since the likelihood of the classes should be maximized, but the loss minimized, the negative log-likelihood is used (Jurafsky and Martin 2023, 12–13).

I decided to use Logistic Regression due to its straightforward applicability (by combining 81 of Scikit-Learn LogisticRegression classifiers with the MultiOutputClassifier[3]) and because it is computationally efficient. In their paper from 2007, Genkin, Lewis, and Madigan (2007) report efficiency as a problem of Logistic Regression on datasets with a large number of predictor variables. However, for the classification task in this thesis, the number of features does not necessarily have to be reduced for an efficient performance. Comparing Logistic Regression to other, non-linear models, Shah et al. (2020, 12–13) report that it outperforms KNN and Random Forest on a small dataset (BBC news texts) in almost all classes

[3] https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html

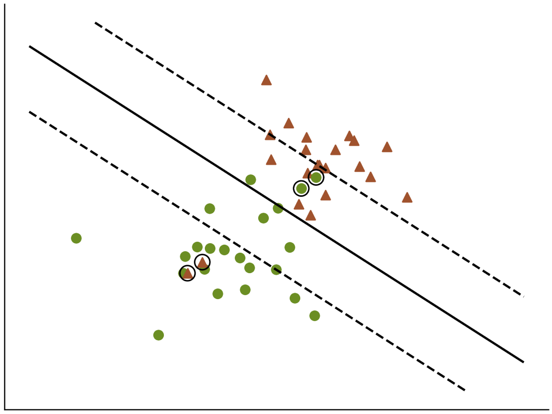


Figure 4.7: Linear SVM (inspired by Kowsari et al. (2019, 29))

in terms of accuracy and in all classes in terms of the F1-score.

Support Vector Machine (SVM) SVMs are vector space-based algorithms. Their goal is to find a decision boundary between two or more classes. While there are many possible decision boundaries, the boundary found by the SVM should ideally be positioned as far as feasible from any data point within the training data. The decision boundary is determined by a small subset of data points of each class. These points are called support vectors (Manning, Raghavan, and Schütze 2009, 119–20). The distance between the support vectors of each class is referred to as the margin, the area between the classes is called the decision hyperplane. Similar to Logistic Regression, the SVM predicts negative or positive scores for each instance. The loss function is called hinge loss, which is 0 as soon as the maximal margin between the support vectors is calculated. Any other data points but the support vectors do not have an impact on the decision boundary. Figure 4.7 shows the separation of two classes by a decision boundary. The dashed lines illustrate the outer boundaries of the decision hyperplane.

If the data are not linearly separable, instances violating the decision boundary can either be penalized with loss terms (in Figure 4.7 those data points that lie within or beyond the decision hyperplane), or non-linearity can be introduced by kernel tricks (Wang and Lin 2015, 187). The hyperparameter C defines how to best separate non-linear problems. A high C parameter creates a smaller margin with both classes close to each other, but there are few instances beyond it. A low C parameter creates a large margin, but there are more instances beyond. The kernel trick, on the other hand, transforms non-linear problems into linear problems. Certain features can be transformed or added or new support vectors can be provided, making the data points better linearly separable. Figure 4.8 shows the separation of the same example data points with a sigmoid kernel.

SVMs are binary classifiers, computing decision boundaries only between two classes. To perform multi-class classifications, it is necessary to either combine

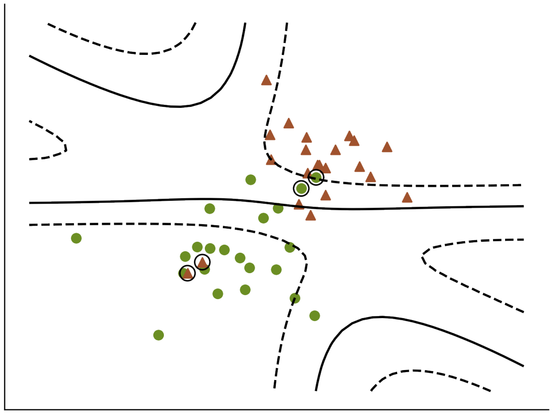


Figure 4.8: Non-linear SVM with Sigmoid Kernel (inspired by Kowsari et al. (2019, 29))

several binary classifiers or use an extended margin concept (Wang and Lin 2015, 187). Combining several binary SVMs predicting one class against all others can be achieved by using Scikit-Learn's MultiOutputClassifier module (Pedregosa et al. 2023).

One of the key advantages of SVMs is their ease of application and that, compared to Logistic Regression, kernel transformations allow the model to deal with non-linear decision boundaries. However, one drawback of SVMs is that the scores generated by this model can be more challenging to interpret than probabilities. In terms of performance, Ikonomakis, Kotsiantis, and Tampakas (2005, 6) report that the precision is usually higher than the recall.

Vinciarelli (2005, 1884) tested several classifiers with artificial noise from stimulated OCR/HTR errors and found that the results achieved by the SVM significantly depend on the noise. In the context of this thesis project, the impact of noise on the scores for the test data is less of an issue as the model will be trained and tested on proofread HTR transcriptions. However, for obtaining and analysing keywords on the unknown data, it has to be kept in mind that the scores evaluated on the test data might not be representative.

As an example of a prior, comparable research project, the study of Yu (2008) can be mentioned. The author applied text classifiers based on Naïve Bayes and SVMs on literary texts (Dickingson's poems and early American novels). Their main finding was that the classifiers are highly domain-specific and not generalizable across texts. While the trained Naïve Bayes classifier could achieve a high accuracy in retrieving eroticism in the poems used as training data, on unknown data its performance decreased (Yu 2008, 341). The task itself is interesting because of its thematic proximity to the digital humanities and the topic of the master's thesis, but it is essential to approach the results with caution, considering that the paper was published 15 years ago.

Neural Network The general idea of neural networks is to simulate biological systems like neurons in the human brain. The previously introduced Perceptron is the most basic architecture for a neural network. It contains a set of input nodes (the feature values) which pass data without undergoing computation. An output node calculates a function of its inputs using the feature weights, similar to synaptic strengths in biological systems (Aggarwal 2015a, 14).

In practice, this very basic architecture is extended by multiple hidden layers between the input and the output layer. Each of the layers in a network takes one representation of the data as input and returns another representation (such as a vector, a matrix, or a tensor) by passing it through functions with iteratively improved weights. In each layer, different features are learned (not necessarily the tokens of the input sequence). Also, the size of the input of one layer does not necessarily have to match the size of the output. In various training epochs, the layers are optimized (since local optimization of a certain layer does not necessarily correspond to the overall, global optimization, one epoch can always be worse than the epoch before).

In the following experiments, I will explore various combinations of layers. In each of these layers, the weights are continuously updated. This happens in two phases: the forward phase and the backpropagation phase. In the forward phase, the weights in all layers are updated and a final output is computed. Comparing the final output with the input's class labels yields an error value. Based on the error value of the forward phase, in the backpropagation phase, the model learns the gradient of one layer as a function of errors and weights in the subsequent layer to compute how much each weight contributed to the overall error. Then the weights are adjusted again, and this adjustment is controlled by the learning rate (Aggarwal 2015a, 15–16).

The neural network built for classifying the Oberfrohna correspondence comprises the following layers:

<u>Embedding Layer:</u>
The input of the embedding layer is a vector of the input document. The words of the input documents are represented by unique word IDs. For each of these word IDs, the embedding layer looks up the corresponding word vector. The word vectors can be either initialized randomly and iteratively optimized during training or be pre-trained (for instance GloVe[4]). For the experiments in this thesis project, no pre-trained vectors will be used. The output of the embedding layer is a matrix (dimension of the vectors × number of words).

<u>CNN Layer:</u>
The convolutional neural network applied in this layer was originally developed for image processing. It analyses segments of an input (originally an image, but the network has also been adapted to textual sequences) and aims to detect certain patterns by moving a set of trainable filters over the input. While images are usually three-dimensional (a tensor: length × width × colour channels), textual

[4] https://nlp.stanford.edu/projects/glove/

input is two-dimensional (a matrix: length of the sequence in tokens × embedding dimension). The width of the filters (how many tokens it considers at once) is a parameter that has to be defined when building the model. Its height is the embedding dimension, which represents the size of the word vectors. The filters are moved stepwise along the input sequence with a certain pre-defined step size (stride). By padding the input sequence by adding zeros to the left and the right (zero-padding), the input dimensions can match the output dimensions and the filter can be moved much further to the boundary of the sequence. Each filter can detect a certain pattern based on learned weights. At each position of the filter, the filter weights and the feature values are multiplied. The sum of the multiplications forms the prediction for that position, activating specific features based on the learned weights. To do so, the ReLU (Rectified Linear Unit) activation function is commonly applied to leave values greater than 0 unchanged and set values less or equal to 0 to 0.

Pooling Layer:
The pooling layer summarizes the result of the previous convolution layer. Average pooling takes only the average value of each filter; max pooling takes the maximum value. Hence, the output of the pooling layer matches the number of filters in the CNN layer.

Bidirectional LSTM Layer:
The LSTM (Long Short-Term Memory) is an extension of the basic Recurrent Neural Network (RNN) architecture. RNNs effectively combine the word vectors learned in the previous layers and can retain only relevant features and suppress insignificant ones. This is achieved by computing a representation (hidden state) for each position in the sequence. The representation of each position is recursively computed from both the total information of all previous positions in the sequence and the information of the particular position (to compute the initial representation in a sequence, the vector preceding the first word is typically initialized to zero). A non-linear function with iteratively optimized feature weights is then applied to the recursively computed word representations (e.g., sigmoid). Instead of one RNN computing the final state from left to right, two RNNs are combined for a bidirectional RNN, computing an output each from left to right and from right to left. The overall output is then combined into one vector. This has the advantage that the information from the beginning of both sides of the sequence is preserved and no information is lost over long distances. The LSTM, the extension of the RNN, was introduced by Hochreiter and Schmidhuber (1997). The problem with the backpropagation in RNNs is that the backpropagated error might either explode (resulting in oscillating weights) or vanish (resulting in unreasonable training time or no effect at all). The LSTM uses a gradient-based algorithm that prevents these issues by enforcing a constant error flow through all layers (Hochreiter and Schmidhuber 1997, 1735–36). Overall, this facilitates the optimization of the parameters.

Pooling Layer:
The global average pooling layer summarizes the information from both the forward and the backward direction of the bidirectional LSTM.

Dense Layer:
In the Dense layer, a non-linear function is applied to the output of the previous layers. In the model architecture applied for this thesis, this is the layer which should provide an output that can be interpreted as probabilities. Therefore, the non-linear function is the sigmoid function, just as for the Logistic Regression model. The output is transformed into a probability distribution across the output classes.

The weights in the layers are initialized randomly. To ensure the reproducibility of the model's results, seeds have to be set for training. Then, the model is trained 10 times with the controlled seeds. This ensures that there is only a 10% probability of being able to train a model with better performance using a seed that is smaller than the highest seed in the set of 10 and higher than the lowest seed in the set of 10 seeds. In the results (Chapter 5.2.4), the performance of the model with the best seed is reported.

Overall, training the neural network with the data of the Oberfrohna correspondence should be seen as an experiment to evaluate if it is even possible to outperform the results of the SVM and Logistic Regression. Deep neural networks trained using conventional methods can generally be expected to exhibit worse performance on small datasets compared to traditional learning algorithms. However, fine-tuned and pre-trained neural networks might have the potential to perform better, even on small datasets (Feng, Zhou, and Hongbiao 2019, 300). Comparing neural networks with three or more hidden layers to an SVM on an even smaller dataset than the Oberfrohna correspondence (487 instances, 2 classes), Feng, Zhou, and Hongbiao (2019, 303, 305) report that the pre-trained neural network could indeed outperform the SVM (accuracy of 0.89 vs. 0.93). However, it is important to keep in mind that this comparison involves a binary classification problem, whereas the Oberfrohna correspondence dataset requires the distinction of 81 classes.

As an example of a research project in a similar field, Chalkidis et al. (2019) tried to classify 57,000 EU legislative documents from the EUR-LEX portal, however on a much larger scale than in the AiK project, with 4300 different labels. Similar to the letters in AiK, many of these labels are very rare. The model architecture that performed best was a BiGRU (another extension of an RNN) with label-wise attention, although using the recurrent neural networks was problematic due to their high computational requirements (Chalkidis et al. 2019, 1, 5).

For the sake of completeness, also transformer models should be mentioned at this point. The previously introduced linear models (Logistic Regression and SVM) have the advantage that they are easy to use on CPU and computationally efficient, while the neural network with recurrent and convolutional layers is more computationally expensive. State-of-the-art models nowadays are often based on transformer architectures. The model architecture was first introduced by Vaswani et al. (2017) and completely relies on an attention mechanism instead of recurrence. HuggingFace Transformers provides a variety of pre-trained transformer-based models.[5] However, an essential disadvantage of all pre-trained models for German texts is that they have a maximum token limit of 512, with padding for longer texts, which is not appropriate for the letters of the Oberfrohna correspondence. An alternative here would be the Longformer model, which can process more tokens (Beltagy, Peters, and Cohan 2020). Unfortunately, it is not pre-trained on German language and can therefore not be used for the Oberfrohna correspondence. Moreover, using a transformer architecture is again critical given the small dataset size. Ezen-Can (2020, 5, 9) compared the performance of the Transformer model BERT to an LSTM on a dataset that is small for the specific task (23,700 instances and 150 classes). The results show that the simplest LSTM architecture worked best on this dataset and could statistically significantly outperform BERT.

4.2.4 Evaluation

The choice of the performance measure depends on the individual classification problem. For the multi-label classification problem with class imbalances across the letters of the Oberfrohna correspondence, the F1-score (also known as F-measure) was selected. It is applied for early stopping the training process using a neural network and predicting with the best epoch, to test different hyperparameters on the development data, and to test the overall performance on the test data.

In general, the insights offered by various performance measures depend on the specific aspects of the classification task under consideration (Kowsari et al. 2019, 45). An alternative approach to measure a classifier's prediction quality would be, for instance, accuracy. Accuracy is a performance measure predicting the proportion of all correctly predicted instances. However, for underrepresented classes, this is hardly suitable. According to Manning, Raghavan, and Schütze (2009, 155), in some binary classification tasks, more than 99.9% of all documents belong to the nonrelevant class. In this case, the model could maximize the accuracy by simply always predicting the negative class. In the labelled documents of the Oberfrohna correspondence, only 0.38% respectively 0.69% of all documents belong to the two smallest classes "Allianzen" ("alliances") and "Ausweise" ("identifications"), which would guarantee high accuracy for these classes if they were simply not predicted at all.

The F1-score, on the other hand, balances the trade-offs among the crucial metrics of true positives, false positives and false negatives, and is able to take class imbalances into account (Leung 2022). These values can be displayed in a confusion matrix:

All instances of the dataset which do have the positive label (TP + FN) are called *relevant*, whereas all instances the model classifies as positive (TP + FP) are called *retrieved*. Using the count of relevant and retrieved instances, the measures of precision and recall can be computed. Precision is the proportion of all retrieved instances that are in fact positive, giving information about the reliability of the

[5] https://huggingface.co/models

		Gold Labels	
		Positive (relevant)	Negative
Predictions	Positive (retrieved)	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 4.1: Confusion Matrix

model:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is the coverage of all relevant instances, giving information about how good a classifier is in finding the relevant instances:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In the one-versus-rest training process of this thesis, a total of 81 different models will be trained, resulting in the creation of 81 distinct confusion matrices.

Since precision and recall give different information about the model, one metric might be more important than the other in certain circumstances (Manning, Raghavan, and Schütze 2009, 165). For instance, for rare categories, recall is the most sensitive measure in terms of how well the classifier performs (Yang 1999, 78). For measuring the performance of the classifiers, the class to which an instance belongs can be determined by a threshold of e.g., 0.5. Depending on the precision and recall on the development data, this threshold can be adapted and might help to achieve better scores (Malato 2021).

The F1-score is a trade-off between precision and recall (Manning, Raghavan, and Schütze 2009, 165). It is appropriate for datasets with imbalanced classes and a small number of class instances compared to the overall number of documents (Yu 2008, 78). Using precision and recall, the F1-score is calculated as:

$$\text{F1-Score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

For multi-class/multi-label problems, the F1-score itself can be divided into three different types: macro, micro, and weighted average (as explained in Leung (2022)):

The **macro average** is the (unweighted) arithmetic mean of the F1-score over all classes:

$$\text{macro-avg} = \frac{\text{F1-score class 1} + \text{F1-score class 2} + \ldots}{\text{number of all classes}}$$

It considers all classes equal and is, therefore, less appropriate for imbalanced datasets such as the Oberfrohna correspondence.

The **weighted average** calculates the F1-score for each class, multiplies these values by the corresponding support proportion (proportion of all instances belonging to this class) and sums up the multiplied values:

$$\begin{aligned}\text{weighted-avg} = {} & \text{F1-score class 1} \cdot \text{support proportion class 1} \cdot \\ & \text{F1-score class 2} \cdot \text{support proportion class 2} + \ldots\end{aligned}$$

where:

$$\text{support proportion for class} = \frac{\text{number of instances for class}}{\text{number of all instances}}$$

The **micro average** is a global average, using the sums of all TP, FP, and FN over all classes to calculate precision and recall and then the F1-score:

$$\text{micro-precision} = \frac{\text{sum of all TP over all classes}}{\text{sum of all TP over all classes} + \text{sum of all FP over all classes}}$$

$$\text{micro-recall} = \frac{\text{sum of all TP over all classes}}{\text{sum of all TP over all classes} + \text{sum of all FN over all classes}}$$

$$\text{micro-avg} = \frac{2 \cdot \text{micro-precision} \cdot \text{micro-recall}}{\text{micro-precision} + \text{micro-recall}}$$

Micro and weighted average are less influenced by a performance that is poorer on the smaller classes than on the larger classes (Leung 2022).

The performance of the models will be reported using the micro average. Since the micro average is calculated by the overall sum of TP, FP, and FP over all classes, it is comparable to the overall accuracy of the multi-class classification problem. It provides information about the general model performance regardless of an individual class (Leung 2022). If required, in the analysis part of the thesis the scores for the individual classes will be reported. In such a highly imbalanced dataset as the Oberfrohna correspondence, the scores for the individual classes can differ vastly. Precision, recall, and F1-score will always be reported within the range of 0 to 1.

Despite being the most appropriate measure, there are still some problems with the F1-score to evaluate the models' performance. Due to the noise in the labels, achieving a high performance on the test data is virtually impossible. Noisy labels of the training data influence the training process, and the noisy labels of the test data make it harder to evaluate the trained model. To compensate for the noisy test data, in addition to the F1-score a qualitative evaluation is required (by closely reading relevant letters). For instance, 826 out of all 1,309 documents are annotated with the keyword "Gefühle" ("feelings"). Since Hilde and Roland's letters are primarily love letters, it can be assumed that most of them deal with feelings—even many of those not explicitly annotated with this keyword. Chances are high that documents counted as false positives with the label "Gefühle" are in fact true positives, however, their gold label is incorrect. Moreover, the summarization process of the old keywords to the new concept terms was crude and it is possible that in the test data some documents are incorrectly labelled with a keyword. In this case, an indeed true negative instance could be counted as a false negative by the performance measure because it was labelled falsely. For further evaluation, also the individual models' performance on the different classes can be examined. Examining the F1-score for each class instead of an average gives information about classes for which a model performs better or poorer. Moreover, it has to be highlighted once more that a score evaluated on the test data might not be representative of the model's performance on the unknown data: On the one hand, when predicting keywords on letters from after 1943, many of these still contain HTR errors, causing additional noise that was not taken into consideration when measuring the performance on the test data. On the other hand, Hilde and Roland’s language and idiolect could have changed over time.

4.3 On the “Humanities” Side: Interpretation of the Keywords

4.3.1 Visualizing the Keywords

The humanities-oriented research questions and hypotheses that were presented at the beginning are dedicated to both the original keywords and the keywords obtained by the text classifier and their interpretation in the historical context. The keywords on all letters over the course of the war must be investigated with a high level of caution and reflection of their reliability. Besides the problem of their unreliability (due to the noisy original labels and the poor results of the prediction), the main issue when investigating timelines of the keywords are gaps in the written correspondence, caused by personal contact of Hilde and Roland before his deployment or during furloughs, the restricted possibility to write during travelling or war captivity, or by letters which were not delivered by the postal service or got lost for any other reason.

For interpreting the keywords, all letters of the correspondence will be aggregated per month. To examine the monthly progression of the individual keywords (indicating months with a relatively higher proportion of letters labelled with a specific keyword compared to others), a line chart is generated for each keyword that exhibits an F1-score on the test data exceeding the score which would result from random class assignment.[6] If required, also graphs for combinations of keywords will be created. The charts visualize the relative share of the letters labelled with the keyword in relation to the total number of letters within a given month. Years would be too extensive periods to compare with each other, while weeks or even smaller units, such as the letters of a single day, are too fine-grained to be meaningful.

The approach has to be used with caution. Only for several months, the correspondence is “complete”, containing as many letters as we can assume that Hilde and Roland would have written under “normal” circumstances of separation in everyday war life (approximately one per day, or even more). Table 4.2 reports the months in which at least 25 letters are available for Hilde and/or for Roland.

For any “incomplete” months it can be assumed that the keywords do not represent Hilde and Roland’s topic focuses appropriately. When plotting the relative keyword frequency over time in a line chart, a keyword can exhibit a substantial relative frequency per month, even if only a single letter or very few

[6]This aims to ensure at least a basic reliability of the frequencies of the examined keywords. The F1-score of the random classifier on the test data for the individual classes serves as a threshold which must be met. Otherwise, they are excluded from further investigation. Applying this threshold, out of the original 81 keywords, only 53 remain. (Out of the 28 removed keywords, 27 had an F1-score of 0 on the test data and one an F1-score slightly lower than when randomly assigned.)

Month	Person	Number of Letters
October 1940	Roland	31
	Hilde	26
December 1940	Roland	27
	Hilde	28
January 1941	Roland	34
	Hilde	29
March 1941	Hilde	32
April 1941	Roland	28
	Hilde	27
May 1941	Hilde	27
July 1941	Roland	25
August 1941	Roland	27
October 1941	Roland	34
	Hilde	30
November 1941	Roland	30
	Hilde	26
December 1941	Roland	26
March 1942	Roland	29
	Hilde	27
May 1942	Roland	25
November 1942	Hilde	32
December 1942	Roland	33
	Hilde	32
January 1943	Roland	31
	Hilde	33
February 1943	Hilde	26
March 1943	Roland	28
May 1943	Roland	31
June 1943	Roland	41
July 1943	Roland	36
	Hilde	29
August 1943	Hilde	30
September 1943	Roland	26
October 1943	Roland	32
	Hilde	28
March 1944	Roland	25
April 1944	Hilde	26
June 1944	Roland	28
July 1944	Roland	26
July 1944	Hilde	27
November 1944	Roland	29
May 1945	Roland	29
July 1945	Roland	27
August 1945	Roland	29
October 1945	Roland	26
November 1945	Roland	28
December 1945	Roland	34

Table 4.2: “Complete” Months (more than 25 letters)

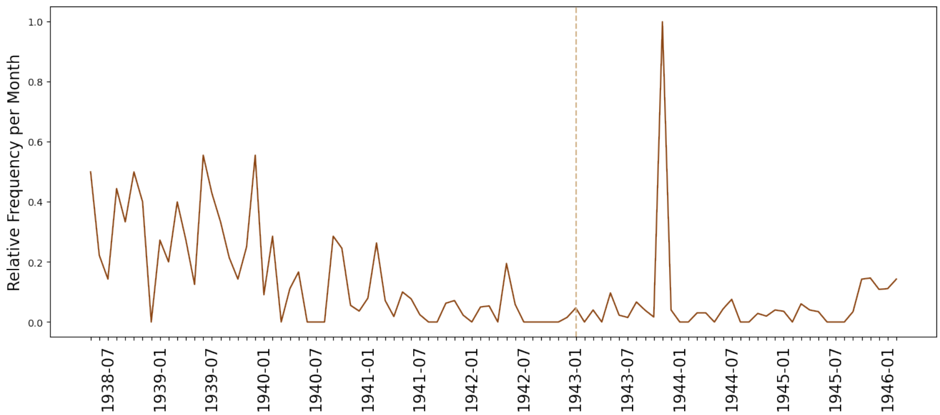


Figure 4.9: “Aus-/Bildung” before Normalization

letters are labelled with it. This can occur when the total number of letters in that particular month is limited. For instance, the keyword “Aus-/Bildung” (“education”) achieves a relative frequency of 1 in November 1944, for which exactly one letter is available, which happened to be labelled with this keyword. Figure 4.9 shows the keyword’s relative frequency. The graph is distorted, and an extreme local peak is created incorrectly (the dashed line indicates in this and all following graphs the beginning of the part of the correspondence where the keywords were assigned automatically by the trained classifier). To avoid distortion caused by the different number of letters per month, the graphs must be normalized. At this stage, we have arrived at the core issue initially posed: the comparability of timelines (see Moeller 2022, 88). As one of the “kreative Ideen und Verständigungsprozesse” (“creative ideas and solutions”) Moeller (2022, 91) calls for, log-transformation is applied to the number of letters per month in the denominator. The relative keyword frequency ($freq$) is calculated by dividing the number of letters labelled with a certain keyword per month (n_{relevant}) by all letters available for this month (n). Applying log-transformation on the denominator, higher values get smaller, which in turn makes the numerator not so small. Smaller numerators, on the other hand, undergo less change. This means that in months with a small number of written letters, the number of relevant letters is divided by a relatively high denominator compared to months with many letters:

$$\text{freq} = \frac{n_{\text{relevant}}}{\log(n)}$$

When using the original relative frequencies before normalization, as in Figure 4.9, there are not only extreme outliers as in November 1944 but also higher relative frequencies of many keywords in the time before Roland’s deployment. In Chapter 4.1, the number of available letters per month was presented. Before Roland’s deployment in July 1940, Hilde and Roland had the opportunity to communicate in person, in addition to exchanging letters. According to our current knowledge,

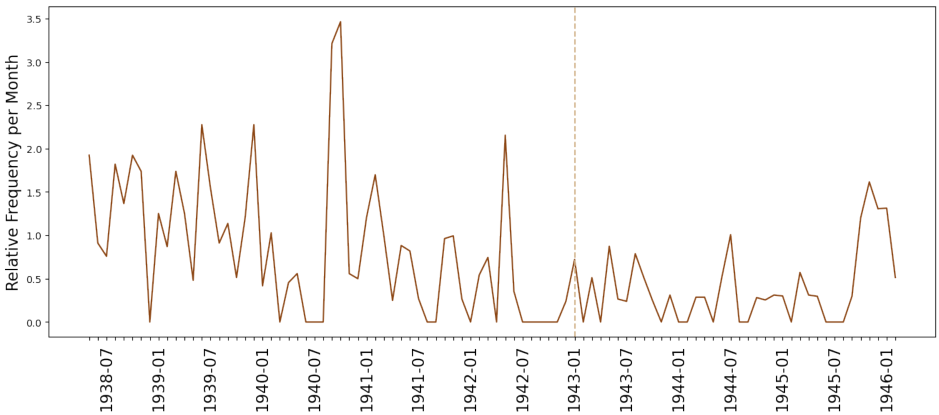


Figure 4.10: “Aus-/Bildung” (normalized)

in this period they wrote an average of 9.44 letters per month. From July 1940 onwards, the corpus contains on average 34.71 letters per month. Naturally, in the months before July 1940, only very few or even only one single letter labelled with a keyword can have a much stronger impact on the relative frequency than in the months later, for which many more letters are available. The single (or limited) occurrence of a keyword per month says relatively little about the actual relevance of a topic for Roland and Hilde. When using log-transformed numbers of letters instead of absolute ones, the relative keyword frequencies in months with fewer letters do not become as high and are better comparable to the relative keyword frequencies in months with many letters.

Since the logarithm of n would be below 1 for the only month in the correspondence with 2 letters (July 1942), it would cause the resulting normalized keyword frequencies for this month to be higher than the numerator n_{relevant} if $n_{\text{relevant}} \geq 0$. The resulting keyword frequencies would be disproportionally high compared to months with 3 or more letters. The logarithm of n for the two months with only one available letter (August 1942 and November 1943) would be 0, causing a division error. To avoid these issues, all three months with only one or two letters are excluded from the plots and their relative keyword frequency is always set to 0. This approach is justified since the calculated frequencies are unlikely to yield significant insights into topic concentration due to the exceedingly low total number of letters for these months.

Using the log-transformed number of letters instead of the actual one can smooth out the created graphs and make the time before and after July 1940 better comparable. For Figure 4.10, the normalized keyword frequencies were used. Even though the frequencies before July 1940 still appear higher, they could be evened out and visually appear not as prominent as previously. Still, one must not look at the graphs without reflection and question irregularities such as this peak. Moeller (2022, 97) refers to Fürber (2016, 28–32, 74–75) who claims that when Distant Reading approaches are pursued, in general variance and room for interpretation must be accepted since it is difficult to estimate error rates unambiguously. Reliability and validity are never guaranteed beyond doubt. In the Oberfrohna correspondence, its incompleteness is an essential part of the corpus. Normalizing the frequencies by log-transforming them is one attempt to ensure a better comparability of frequencies over time.

For analysing the rise and fall of certain topics over time, the created plots are studied for anomalies (for instance conspicuously high points in the graph). During significant and intriguing periods, the actual letters are subsequently referenced and subjected to more thorough scrutiny. This process aims to ascertain whether the hypotheses hold true or are more likely to be disproven.

4.3.2 Similar Research

In general, the method used in this master's thesis (using a text classifier to assign keywords to a complete corpus, and based on this, analysing the progression of keywords) seems novel. However, there is research literature on comparable issues (the change of topics and discourses in large corpora over a certain period). The documents are then mostly assigned to the topics of the corpus by a topic model. Topic modelling is a very popular method in the field of the digital humanities. Due to its research objectives which are quite often similar to those in the thesis, it should be briefly introduced.

Topic modelling is a statistical tool that obtains probability distributions of words and corresponding topics (Au Yeung and Jatowt 2011, 3). It should give researchers an idea of which particular themes are or are not addressed in a document (Fridlung and Brauer 2013, 152). Even though obtained by a different method, the keywords assigned by the text classifiers can be used in a similar way as topics from a topic model for the following interpretations. There are various studies applying topic modelling not only in literary studies but also in history. Often, topic models are applied to newspapers in order to study variances in discourses over time. Despite the difference in the methods of how the topics are obtained, their use and attempts of interpretation can be similar to the research focus of the thesis, which aims to study variances of topics over time: Au Yeung and Jatowt (2011) apply a topic model to understand modifications and eradications of collective memories using news articles. Marjanen et al. (2020, 2) use the example topics of church/religion and education to study discursive changes in Finish media. Viola and Verheul (2020, 912) study public discourses in Italian ethnic newspapers over the course of 22 years. Lee (2019) combines critical discourse analysis, a qualitative method, with topic modelling to study the media representation of immigrant workers in Korean news reports. Pursuing a similar goal but with a different kind of data, Indukaev (2020) uses internet texts from a Russian media database to study ideas in Russian politics.

Although topic modeling is one of the most popular methods in digital history nowadays, Fridlung and Brauer (2013, 156) noted that (at least until 2012) it is only seldom used for actual exhaustive and in-depth historical analyses. Instead of being dedicated to a specific research question, most papers would focus on the method itself and its applications. To avoid this risk, precise research questions that target explicit topics are required. These were formulated at the outset in the form of the research questions and hypotheses.

Chapter 5

Data Analysis: Predicting and Interpreting Keywords

5.1 Data Preparation

The first step of working with the data of the Oberfrohna correspondence involves creating a complete, comprehensive dataset. To accomplish this, I gathered all available letters from both the T&S WordPress blog and Transkribus, subsequently combining them into a unified CSV file that encompasses the entirety of the correspondence. This file, preserving the data and metadata of the letters, is structured with the following columns:

- *AID* (author ID: 1 = Roland, 2 = Hilde)
- *LID* (letter ID: number of the letter written by this author on this day)
- *LetterSignature* (OBF-yymmdd-00AID-LID: OBF (Oberfrohna) as a unique identifier for the Oberfrohna correspondence on the AiK website bringing together three correspondences; the date; the author ID preceded by two zeros;[1] and the letter ID as a two-digit integer[2])
- *LetterDate*
- *Text* (the actual letter)
- *Keywords* (a list of strings, if available)

The first half of the raw data preserved in the file originates from the former project website, the WordPress blog of T&S, and covers all letters until March 1942. They were exported from T&S WordPress rather than AiK because I pre-processed them in December 2022, when the AiK website was not online yet and T&S still existed. WordPress content can be exported in XML. All letters as well as the corresponding metadata could be extracted from the XML files. Further data cleaning included deleting all annotations, hyperlinks, and in-text comments made by bloggers. When bloggers added images in T&S, they usually wrote a short description contextualizing the image. Image descriptions are unfortunately

[1] The correspondence also contains very few letters written by another person than Roland and Hilde. For now, these people are captured by the author ID 3. Using a three-digit integer for the author in the letter signature ensures that in case many more letters by different authors are found, all of them can be represented by an appropriate ID without having to change the format of the signatures.

[2] Representing the letter ID as a two-digit integer should ensure that all letters written on one day can be captured by the Signature without having to change its format. However, to our current knowledge, Hilde and Roland never wrote a two-figure number of letters on one and the same day. In the entire corpus, there are only 13 letters with the letter ID 3 and only 3 with the letter ID 4.

not in between specific XML tags. If bloggers added no specific keyword, such as “Abbildung” (“figure”) in the image descriptions, they could hardly be kept apart from the letter content, and it is still possible that there are some image descriptions in the cleaned data.

The other half of the raw data had to be exported from Transkribus and pre-processed afterwards. Transcribed text can be exported with annotations in XML format from Transkribus. The main issue here was that letters are uploaded and transcribed month-wise in the AiK project. All pages written by either Hilde or Roland in one month are combined into one single PDF before the upload to Transkribus. Their transcripts then contain page break elements (<pb n='[number of page]'/>), but it is not marked where one letter ends and a new one begins. Dates or salutations are too inconsistent to rely on them as clear markers of a letter’s beginning. To address the issue of detecting the final page of one letter and the beginning page of another, a semi-automated approach involves comparing the transcripts with the original scans. All individual scanned pages are numbered consecutively with the date, the author ID, and the number of the letter written by this author on this day (this is the same for all pages of a letter). Moreover, an additional character marks the beginning of an individual page (e.g., 441202-1-1a for the first page of the first letter written by Roland on December 2, 1944). Accordingly, each XML export for one month has the same number of page break elements as there are individual scanned pages for this month. Each scan marked with *a* is the first page of a new letter, and its corresponding page break element in the XML indicates the beginning of a new letter. This allows for the identification of pertinent page break elements, facilitating the division of individual letters when new ones commence on new pages. This strategy cannot be employed in months where the commencement of new letters does not align with new pages. In some of his letters from 1945 and 1946, Roland continues his subsequent letter on the same sheet of paper or page in his notebook if there is available space following the completion of the previous one. Therefore, 16 PDF files consisting of letters from Roland had to be split manually.

After splitting all letters, the metadata (AID, LID, and date) had to be created, and annotations had to be removed. Hilde and Roland’s true names were replaced by the pseudonyms. Frequently, Transkribus transcribed the German letter “sharp s” (*ß*) with the Greek letter beta (β). Subsequently, the incorrect character had to be replaced with the correct typographical character. This adjustment enables smoother processing in a later step such as lemmatization. Most of the letters exported from Transkribus are not proofread twice, many not even once (the letters from September 1943 onwards, which affects around 35% of the entire corpus), which causes some inherent noise in the dataset.

The data which are in a consistent format with their metadata can be further processed to a suitable format to train the classification models. While each applied model needs its specific input format, which will be explained in the corresponding subchapters, certain pre-processing steps are similar for some of them. Tokenization, stop word removal, lowercasing, and further cleaning of noise and unnecessary features are very basic pre-processing steps that many statistical and probabilistic learning algorithms can benefit from (Kowsari et al. 2019, 4).

The spaCy lemmatizer[3] (Montani et al. 2020) tokenizes (splits into lists of separate tokens) and lemmatizes the input text. It aims to convert tokens to their base form using a lookup table. It analyses the words morphologically and, if possible, returns the lemma (dictionary form) instead of the word. The main advantage of lemmatizing is feature reduction, as it removes inflected word forms and collapses them all into one lemma (e.g., “gone” → “go” but “going” → “going”). Since German is a highly inflected language, the related pre-processing method stemming is hardly possible. Opposed to lemmatizing, stemming uses a heuristic process to cut off the inflected suffixes (e.g., “going” → “go” but “gone” → “gone”). With stemming, the majority of related forms of a word can be consolidated, while lemmatization collapses various inflectional forms into a single base form. Stemming is common for English text data, however not appropriate for a language with more morphology, such as German (Manning, Raghavan, and Schütze 2009, 32–33).

For stop word removal, I used the German stop word list from spaCy.[4] With a length of 543, the spaCy stop word list is already quite extensive. To adapt it to the corpus, I extended it with additional words, such as German weekdays and months. These are highly frequent in the corpus, as Hilde and Roland typically commenced their letters with the present date and frequently spelt out both the day and month. For a classification model, they do not bear any semantic content and are completely randomly distributed over all classes. The main purpose of a stop word list is to reduce the number of features that have to be stored (Manning, Raghavan, and Schütze 2009, 27). But in the letters, the weekdays and month names might not only be unnecessary features, but a model could treat them as actual signals for certain classes. More stop words added were words in old German spelling using “sharp s” (ß) instead of “ss” (“muß”, “daß”) and some of Hilde and Roland’s dialectal/idiolectal terms (“sooo”, “habn”). The word clouds in Figure 3.6 also show that nicknames the spouses gave each other are some of the words with the highest frequencies. These are mostly not semantically relevant for the corpus either. Therefore, a separate stop word list of nicknames was created to supplement the original stop word list. Both lists can be found in Appendix B.1. At this point, however, the danger of excessive feature reduction must be pointed out. For instance, Yu (2008, 341) identified stop words as highly discriminative features for one of the text types they trained a classifier on. For this project, the results without filtering out any stop words will be compared to the basic stop word list and the one expanded by nicknames.

The third pre-processing task that I applied to the data is lowercasing (all words in the corpus). For English data, it is possible to only lowercase the first word of each sentence, as lowercasing the entire document carries the danger of collapsing words that were supposed to be kept apart (e.g., “Windows” as opposed to “windows”) (Manning, Raghavan, and Schütze 2009, 30). This approach is not possible in German due to the capitalization of nouns and the marginal danger of creating polysemy (e.g., “Weg” as opposed to “weg”) by lowercasing must be

[3] https://spacy.io/api/lemmatizer

[4] https://github.com/explosion/spaCy/blob/master/spacy/lang/de/stop_words.py

taken into account.

For fitting the classifier models, all labelled documents are split into a training (60% of data), development (20%) and evaluation set (20%). Instead of a random split, they are split chronologically, with the earliest letters used as training data and the later ones used as development and test data. This should help to deal with the fact that the themes the spouses wrote about as well as language and vocabulary might have changed over time. With a random split, for instance, letters from 1939 could be used as test data. However, the results on the test data should be as representative as possible for the unlabelled data. When the classifier is applied to the letters written from 1943 to 1946, a performance that was measured on the very early letters might not be representative.

5.2 Training the Models

The results of the text classification models are presented below. Various parameter combinations were tested to achieve the best possible outcome. To train the selected models with the data, the raw data need to be formatted into a suitable input format. Logistic Regression, SVM, and the rule-based classifier all require the lemmatized letters as their input. In the case of Logistic Regression and SVM, also stop words can be removed.

In addition to the issues with the data and their labels that have already been discussed (imbalanced classes, ambiguity, overlaps, and lack of ground-truth, see Chapter 4.2.2), another factor that might influence the performance of the classifiers negatively is the varying number of labels with which the individual letters are annotated. A training document (written between May 1938 and July 1941) is on average annotated with 9.12 labels and therefore does not properly represent a document of the test data (written between May 1942 and December 1942), which is on average annotated with 5.00 labels. Conversely, the average letter length does not exhibit a chronological decrease, which could have correlated with a reduction in the number of labels. Instead, it is noteworthy that the test set has the highest average letter length (see Table 5.1).

Dataset	Mean Tokens per Document	Mean Labels per Document	Documents
Training Set (05/1938-07/1941)	1079.2	9.12	782
Development Set (07/1941-05/1942)	1280.16	6.05	263
Test Set (05/1942-12/1942)	1389.21	5.00	264
Unlabelled Documents (01/1943-03/1946)	1177.89		1322

Table 5.1: Dataset Size

The Scikit-Learn Dummy Classifier[5] serves as an absolute baseline for the text classification results to identify classes on which the models perform worse than random chance. Using the parameter “stratified”, it randomly assigns class labels based on the multinomial distribution of the class probabilities of the training data (Pedregosa et al. 2023). Thus, the number of classes assigned to the test data is always based on the training data. However, the test data on average has fewer labels per letter compared to the training data. Since the classes are randomly selected and there is no danger of overfitting, for the sake of illustration, the random classifier can also be based on the development data. The comparison reveals that the recall (and therefore the micro average) is slightly higher when using the training data, as simply more labels per document are retrieved, while the proportion of how many of them are relevant does not change. The random classifier was trained ten times. The means of the achieved precision, recall, micro average, and number of labels per document are reported in Table 5.2.

	Precision	**Recall**	**Micro Average**	**Labels per Document**	**Unlabelled Documents**
Performance on development data, trained on **training data**	0.144	0.219	0.174	9.20	0
Performance on development data, trained on **development data**	0.156	0.157	0.156	6.05	0

Table 5.2: Scores for a Random Classification

5.2.1 Rule-Based Classifier

The rule-based classifier assigns a document to a class if the corresponding keyword can be found in the lemmatized text. The list of 81 keywords (concept terms) is extended with the 1,689 old keywords that were combined (see Appendix A.1). Some of the new concept terms consist of two words, e.g., “Aus-/Bildung” (training/education). To form the rule, it is divided into the two parts “Ausbildung” and “Bildung”. If either one or both of these words can be found in a document, it will be assigned to the class “Aus-/Bildung”. Also any occurrence of an old keyword which was included in a new keyword forms a rule to assign the new keyword to a letter (e.g., 18 old keywords signify that a letter belongs to the class “Aus-/Bildung”).

[5] https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

One problem with this approach is that the performance on the test data might not be representative of a performance on the data that have not been labelled so far. The old keywords and partly also the concept terms to which they were summarized originate from the letters that had already been processed. Following the old labelling system, a completely different set of words could be created for the so far unprocessed letters. Such a set of words might produce better results on the unlabelled data than the set of words created using the already labelled data. Another issue with the approach is the noise in the uncorrected HTR transcriptions. Most likely, there are many documents which would match with some keywords if they had already been corrected, but due to HTR errors, occurrences of certain words cannot be identified as matches with the rules. Therefore, the rule-based classifier should be used with caution to label the unlabelled data.

For the rule-based classifier, a high recall and low precision can be expected. Since the total set of all old keywords contains 1,689 words and 1,591 of these were used to create new concept terms, several matches with the lemmatized texts can be expected. For instance, in the process of summarizing labels, a decision was made to include the old label "Marsch" (march) into the concept term "Aus/Bildung". The rule-based classifier will assign any document with an occurrence of "Marsch" to this class. However, the mere occurrence of this word may not indicate that a letter necessarily discusses the topic "Aus-/Bildung". Similar to this example, many letters could be assigned to incorrect classes.

Because the rule-based classifier operates by applying the predefined rules on all documents without the need for training data, the performance can be reported and compared on all three splits of the entire dataset (Table 5.3).

	Precision	**Recall**	**Micro Average**	**Labels per Document**	**Unlabelled Documents**
Training Set	0.307	0.744	0.435	22,07	1
Dev Set	0.184	0.703	0.292	23,02	0
Test Set	0.180	0.702	0.287	23.30	0

Table 5.3: Scores for the Rule-Based Classifier

The notable discrepancy in performance (particularly concerning precision) between the training set and the development or test set can likely be attributed to the substantial difference in their sizes, with the training set being three times larger. The rules were created based on the available data. With a 60:20:20 split, 60% of the rules were created by the training set and only 20% each based on the development and test sets. The results of the rule-based classifier are therefore also a representation of how the rules were created.

Moreover, a major difference in results between precision and recall can be observed. As could be expected, the set of in total 1,774 rules matches a large number of false positive documents. Compared to the gold annotations, for each dataset, there are between 2.5 and 4 times more labels. Using a stemmed list of rules (for instance "ungewiss" ("uncertain") instead of "ungewissheit" ("uncertainty")) proved to be inappropriate as it even increased the recall, but the corresponding decrease in precision was not justifiable.

The individual F1-scores on the test data per keyword range from 0,66 (“Gefühle”) to 0 (“Allianzen”, “Antisemitismus”, “Ausweise”, “Baukunst”, “Bevölkerung”, “Gräueltaten”, “Sprache”, “Status”). Table 5.4 shows the scores on the test data of the 20 most common keywords in the entire labelled data.

Keyword	Precision	Recall	Micro Average	Occurrence in entire labelled data
Arbeit	0,215	0,878	0,345	272
Begegnungen	0,078	0,545	0,136	242
Familie	0,245	0,963	0,391	498
Feste	0,32	0,904	0,473	328
Gefühle	0,496	0,984	0,66	826
Geschlechterrollen	0,333	0,032	0,059	201
Gesundheit	0,232	0,821	0,362	207
Glaube	0,227	1	0,369	428
Hausarbeit	0,299	0,64	0,408	173
Kameraden	0,258	0,852	0,397	179
Kinder	0,218	0,633	0,325	176
Kriegsverlauf	0,273	0,891	0,418	262
Mobilität	0,142	0,815	0,242	319
Paarbeziehung	0,124	0,481	0,197	181
Schreiben	0,362	0,872	0,512	476
Werte	0,11	0,81	0,194	167
Wetter	0,292	0,568	0,385	312
Wirtschaft	0,294	0,689	0,412	282
Zeit	0,143	0,771	0,241	201
Zukunft	0,14	0,71	0,234	261

Table 5.4: Results of the Rule-Based Classifier (20 most common keywords)

While “Gefühle” can be predicted by 117 different rules, for the classes with a score of 0 there are between 4 and 11 rules. In general, a tendency of an increasing F1-Score with an increasing number of rules can be observed (Figure 5.1).

Having a large number of rules for one label may result in a high number of false positives for that label. However, upon closer analysis of the false positives, it turns out that they are not only caused by a large number of rules, but also by misleading, incorrect, or overly general rules (Figure 5.2). The highest number of false positives was obtained for the class “Lebenszyklus” (“life cycle”), which can be obtained by only 9 different rules. It comprises the terms “Geburt” (“birth”), “Laufbahn” (“career”), “Leben (“life/living”), “Schwangerschaft” (“pregnancy”), “Sterben” (“dying”), “Tod” (“death”) and “tot” (“dead”). In the word clouds with removed stop words and nicknames (Figure 3.7), the importance of the word “Leben”/”leben” in Hilde and Roland’s letters is visible. Apparently, “Leben” is a misleading rule for the rule-based classifier. Removing it from the list of rules, the F1-score for the class “Lebenszyklus” increases from 0.09 to at least 0.13, decreasing the number of false positives from 207 to 47. The second highest number of false

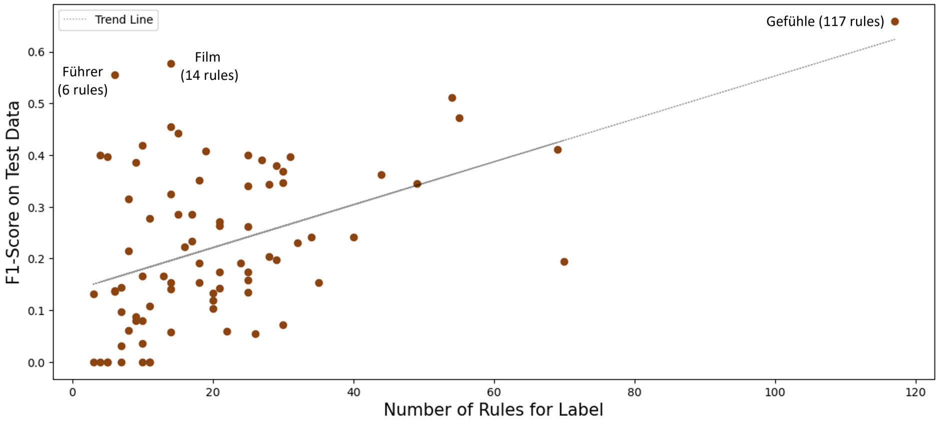


Figure 5.1: Correlation Rules—F1-Score

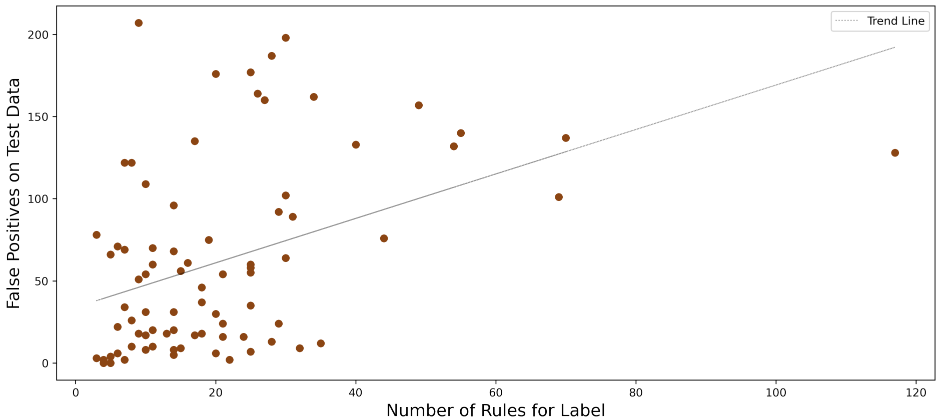


Figure 5.2: Correlation Rules—False Positives

positives has the class “Glaube” (“belief”) due to the rule “Gott” (“God”), which is a frequent word in pious Hilde and Roland’s usage. Similar to these examples, manually refining rules results in avoiding some false positives and would enhance the precision of the rule-based classifier. Also, this indicates that the rule-based classifier still has a potential to improve when using a more carefully and manually curated set of rules. The misleading rules “Leben” and “Gott” will be removed for training the stacked classifier.

5.2.2 Logistic Regression

The Logistic Regression classifier is built with Scikit-Learn's Logistic Regression module (Pedregosa et al. 2023).[6] For the input, a design matrix X (documents × features) and a label matrix y (documents × classes) are required. A bag-of-words representation of the documents serves as a feature extraction method. In the bag-of-word (BOW) representation, all unique single words found in the training set (apart from the stop words) are represented as features. Their feature values are the count of their occurrence in a document. As a result, the natural structures of the text are broken and semantic relationships between words get ignored, making the text understandable by the algorithm but less so by human readers (Scott and Matwin 1999, 3–5). The main challenge of BOW representations usually lies in their scalability, as vocabularies can get very large (Kowsari et al. 2019, 6). Lemmatizing/stemming and stop word removals are approaches to deal with this problem (Scott and Matwin 1999, 3). Moreover, the words which are used as features can be narrowed down by considering for instance only the n) most frequent words, or excluding words with a higher or lower document frequency than a pre-defined threshold (Pedregosa et al. 2023). For training the regression model as well as the SVM, there were no performance issues when using the complete vocabulary.

For the experiments, the vocabulary including all stop words, the vocabulary after basic stop word removal, and the vocabulary after extended stop word removal were used to create the BOW. In terms of efficiency as well as performance, there is hardly any difference. However, in several experiments, binary features, which further reduce the dimensionality, yielded a better performance than word counts. In general, feature counts per document can help to "determine the focus points of the documents" (Kowsari et al. 2019, 6), but in the Oberfrohna correspondence, they seem to have a negative impact on the performance. One possible reason might be that all relevant information about a document belonging to a class is already contained in the very fact that a certain word occurs. Different feature values defined by the count of occurrence mean that not the plain occurrences of a word are considered, but that they are also weighted. In a (hypothetical) letter assigned to the classes "Machthaber" ("rulers") and "Glaube" ("belief"), "Hitler" is mentioned perhaps only once, while "Gott" ("God") can have a much higher feature value, but both are equally relevant for the topics of the letter. An additional feature that turned out to be useful for the representation of the documents is the author (either Roland or Hilde).

Overall, the one-hot-encoded document-term-matrices used for Logistic Regression as well as for the SVM, have the shape n × (m+2), where n = number of documents in training, development, or test set, and m = number of unique words in training set + letter is written by Hilde + letter is written by Roland. The output is a matrix of the shape n × k, where k = 81 (number of classes and their corresponding keywords in the corpus). Since the Logistic Regression returns probabilities instead of absolute values, these are converted to absolute values by Scikit-Learn's Logistic Regression module with a threshold value of 0.5.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

The Logistic Regression was fine-tuned with the following parameters:

- Regularization: L1
- C (inverse of regularization strength; the smaller, the stronger): 0.2
- Solver (optimization algorithm): liblinear (appropriate for smaller datasets with L1 regularization)

Using this best combination of parameters which was determined by a grid search, the scores reported in Table 5.5 can be achieved on development and test data, with different levels of stop word removal.

	Prec.	**Rec.**	**Micro Average**	**Labels per Doc**	**Unlabelled Docs**
Dev Set with Stop Words	0.267	0.459	0.338	10.40	0
Test Set with Stop Words	0.254	0.463	0.328	10.92	0
Dev without basic Stop Words	0.261	0.493	0.342	11.40	0
Test Set without basic Stop Words	0.241	0.464	0.317	11.57	0
Dev Set without extended Stop Words	0.260	0.484	0.338	11,25	0
Test Set without extended Stop Words	0.239	0.457	0.314	11,47	0

Table 5.5: Scores for the Logistic Regression

On average, Logistic Regression predicts more labels for each document than the labels that are originally assigned, but not as many as the rule-based classifier. The correlation coefficients Spearman's Rho of 0.6 and Pearson's R of 0.61 suggest a high correlation between how often a class is represented in the training data and how well it is predicted on the test data. Still, there are some outliers, as can be seen in Figure 5.3. Even though the classes "Führer" and "Film" ("film") are relatively infrequent, the classifier learns to a certain degree how to predict them. On the other hand, for "Begegnungen" ("encounters") the Logistic Regression performs worse than for other keywords with a comparable amount of training documents.

For the 20 most common keywords in the entire labelled dataset, the Logistic Regression has the results displayed in Table 5.6 on the test data.

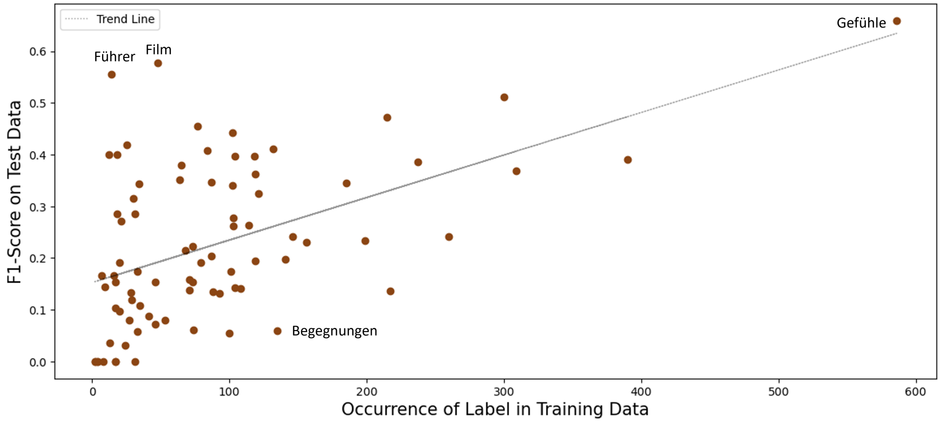


Figure 5.3: Correlation Instances—F1-Score

Keyword	Precision	Recall	Micro Average	Occurrence in entire labelled data
Arbeit	0,254	0,367	0,3	272
Begegnungen	0,036	0,091	0,051	242
Familie	0,336	0,685	0,451	498
Feste	0,5	0,575	0,535	328
Gefühle	0,532	0,852	0,655	826
Geschlechterrollen	0,283	0,419	0,338	201
Gesundheit	0,289	0,393	0,333	207
Glaube	0,311	0,862	0,457	428
Hausarbeit	0,429	0,48	0,453	173
Kameraden	0,246	0,593	0,348	179
Kinder	0,29	0,667	0,404	176
Kriegsverlauf	0,304	0,761	0,435	262
Mobilität	0,2	0,556	0,294	319
Paarbeziehung	0,121	0,296	0,172	181
Schreiben	0,419	0,419	0,419	476
Werte	0,178	0,381	0,242	167
Wetter	0,239	0,568	0,336	312
Wirtschaft	0,468	0,59	0,522	282
Zeit	0,186	0,314	0,234	201
Zukunft	0,172	0,516	0,258	261

Table 5.6: Results of the Logistic Regression (20 most common keywords)

5.2.3 Support Vector Machine

The SVM classifier is also built with a module from Scikit-Learn (Pedregosa et al. 2023).[7] Input and output matrix look the same as for the Logistic Regression. For the SVM, similar to Logistic Regression, the best results could be achieved with binary features and the letter author as an additional feature. The removed stop words influence the results again only slightly. Aggarwal and Zhai (2015, 315) report a "general consensus" that non-linear versions that can be implemented in originally linear classifiers (such as the non-linear sigmoid kernel instead of the linear decision boundary) usually do not pay themselves off, as text data tends to be linearly separable. However, when comparing an SVM with linear decision boundary to an SVM with non-linear kernel functions (in Scikit-Learn, a polynomial kernel, Gaussian RBF kernel, and sigmoid kernel are implemented (Pedregosa et al. 2023)) on the Oberfrohna correspondence, the sigmoid kernel does achieve the best results on both development and test data.

The best regularization strength for the dataset (for both sigmoid and linear kernel) is 0.9. Moreover, Scikit-Learn's LinearSVC module allows to set the class weight to "balanced". This is useful for the imbalanced classes in the Oberfrohna dataset, as the feature weights are adjusted inversely proportional to the classes' frequencies in the label matrix (Pedregosa et al. 2023).

Using linear and sigmoid kernels, the performance reported in Table 5.7 can be achieved.

	Kernel	Prec.	Rec.	Micro Average	Labels per Doc	Unlabelled Docs
Dev Set with Stop Words	Linear	0.351	0.223	0.273	3.84	4
	Sigmoid	0.268	0.377	0.313	8.5	5
Test Set with Stop Words	Linear	0.363	0.232	0.283	3.83	7
	Sigmoid	0.2678	0.390	0.317	8.73	8
Dev Set without basic Stop Words	Linear	0.357	0.236	0.285	4.0	1
	Sigmoid	0.296	0.364	0.326	7.4	8
Test Set without basic Stop Words	Linear	0.351	0.227	0.276	3.89	12
	Sigmoid	0.287	0.358	0.319	7.48	14
Dev Set without extended Stop Words	Linear	0.319	0.285	0.301	5.40	1
	Sigmoid	0.305	0.367	0.333	7.27	10
Test Set without extended Stop Words	Linear	0.310	0.261	0.283	5.04	8
	Sigmoid	0.296	0.358	0.324	7.25	16

Table 5.7: Scores for the SVM

Compared to the sigmoid kernel, the micro average recall is by far lower with a linear decision boundary. As a result, the average number of assigned labels

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

is smaller and there are several unlabelled documents in the test set. However, the precision is slightly higher. Ikonomakis, Kotsiantis, and Tampakas (2005, 6) report poor recall with excellent precision as typical for SVMs in text classification. Even though the precision of 0.31 is by far not “excellent”, compared to any other classifier applied to the Oberfrohna correspondence the linear SVM does have the highest precision. When inspecting the performance of the two SVM models as well as the Logistic Regression on the individual classes, it becomes evident that especially the linear SVM struggles to learn and predict multiple classes effectively (Table 5.8).

	Logistic Regression	**Linear SVM**	**SVM with Sigmoid Kernel**
Number of Classes with F1-Score = 0	19	55	40

Table 5.8: Comparison Logistic Regression—SVM (F1 = 0)

The SVM with a sigmoid kernel fails to predict 40 classes on the test data. For these 40 classes, there are on average 34.4 training documents. For the other 41 classes, for which the SVM with sigmoid kernel can achieve an F1-score > 0, there are on average 140.3 training documents.

Table 5.9 compares the results on the individual classes of the SVM with sigmoid kernel with those of the Logistic Regression and reports the count and proportion of classes in which the SVM is better than the Logistic Regression.

F1-Score in Comparison	**Absolute number of Labels**	**Relative number of labels**
Better than Logistic Regression	22	27.1%
Worse than Logistic Regression	51	63.0%
Equal (both SVM and LR = 0)	8	9.9%
Sum	81	100%

Table 5.9: Comparison Logistic Regression—SVM (F1 per Class)

Despite the 40 classes for which the SVM with sigmoid kernel fails to learn effectively, it can slightly outperform the Logistic Regression, for instance, when comparing the class “Begegnungen” (F1-score 0.09 vs. 0.05). The class was previously pointed out as an outlier for the Logistic Regression (achieving a relatively low F1-score of 0.051 despite the 217 training documents). The class “Führer”, for which the Logistic Regression performs surprisingly well (F1-score 0.556) despite the low number of training documents (14), cannot be predicted at all by the SVM (F1-score 0). Also, for the class “Film” (48 training documents), the Logistic Regression outperforms the SVM (F1-score 0.577 vs. 0.32).

Table 5.10 reports precision, recall, and F1-score that the SVM achieves on the 20 most common classes.

Keyword	Precision	Recall	Micro Average	Occurrence in entire labelled data
Arbeit	0,27	0,204	0,233	272
Begegnungen	0,091	0,091	0,091	242
Familie	0,322	0,519	0,397	498
Feste	0,56	0,384	0,455	328
Gefühle	0,56	0,844	0,673	826
Geschlechterrollen	0,255	0,419	0,317	201
Gesundheit	0,353	0,429	0,387	207
Glaube	0,326	0,81	0,465	428
Hausarbeit	0,41	0,32	0,36	173
Kameraden	0,212	0,519	0,301	179
Kinder	0,348	0,533	0,421	176
Kriegsverlauf	0,298	0,609	0,4	262
Mobilität	0,205	0,296	0,242	319
Paarbeziehung	0,169	0,407	0,239	181
Schreiben	0,432	0,407	0,419	476
Werte	0,192	0,476	0,274	167
Wetter	0,239	0,432	0,308	312
Wirtschaft	0,514	0,623	0,563	282
Zeit	0,083	0,029	0,043	201
Zukunft	0,193	0,71	0,303	261

Table 5.10: Results of the SVM (20 most common keywords)

5.2.4 Neural Network

The neural network is trained with Keras, the API of the library Tensorflow with a focus on machine learning and especially deep learning (Abadi et al. 2015).[8] For training the neural network, the stop words are not removed, so the model can learn the complete sequences of tokens (Gunasekara and Nejadgholi 2018, 23). The dictionary with word IDs is created based on the 100,000 most common words in the training data (words in development and test data that do not occur in the training data do therefore not have an ID and are replaced by a special token representing unknown words). Since the input shape of all sequences (the individual letters) must be consistent to train the model, all sequences are padded (or truncated) on the left. The maximum number of tokens (2,204) is set higher than the average letter length (1,223) to ensure that hardly anything is truncated so as not to lose any information. The letters are structured in a way that different topics are discussed one after the other, and even in the very last paragraph, a new topic can be mentioned which would assign the letter to another, additional class. Only 5% of all letters are longer than 2,204 tokens. This turned out as a suitable maximum token length. Truncating only the 5% of all documents with more than 2,204 tokens does make the training of the model computationally expensive and it will take longer than, for instance, truncating all letters longer than the average

[8] https://www.tensorflow.org/guide/keras

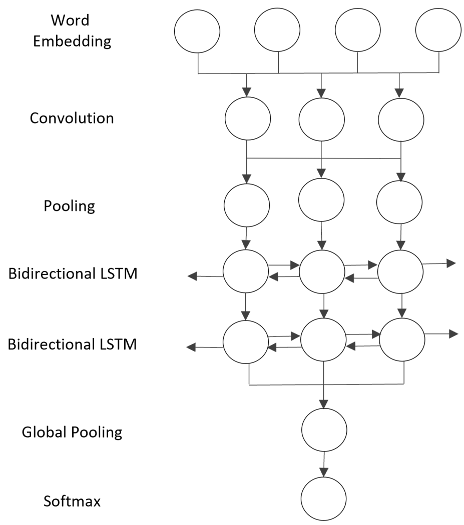


Figure 5.4: Neural Network Architecture

letter length. However, the time spent is still reasonable, compared to padding all letters to the length of the longest letter.

The network in this thesis consists of an embedding layer[9], a convolution layer[10], and two bidirectional LSTMs[11] (Figure 5.4).

The following model parameters were used to fine-tune the network:

- Embedding dimensions (size of the word vectors initialized in the first layer): 50. Most likely, a higher embedding dimension might lead to better results. However, only word embeddings with the length of 50 were applied to save GPU memory for a easier training process.
- Number of filters for the CNN (a common, suitable number is approximately as many as embedding dimensions of the word vectors—however, in the experiments a slightly lower number turned out to work better): 32
- Kernel size (size of the filter = number of positions simultaneously looked at by the filter): 5
- Activation function for CNN: ReLU (common for hidden layers)
- Pool size for the CNN (since max pooling will be applied, the pool size defines the number of positions of which the pooling layer will take the maximum): 3
- Units of the bidirectional LSTM (dimensionality of the LSTM output shape): 16

[9] https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding
[10] https://www.tensorflow.org/api_docs/python/tf/keras/layers/Conv1D
[11] LSTM: https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM; Bidirectional: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Bidirectional

- Optimizer for the entire network: Adam (Kinga and Ba 2015)

The model is trained in various epochs. The early stopping callback[12] allows the training to stop as soon as the micro average does not improve anymore. The patience parameter can be set to keep training for several epochs in case the F1-score does improve again. For the experiments in the thesis, it is set to 3. After training over 10 different seeds, the performance reported in Table 5.11 can be achieved with the best seed.

	Precision	**Recall**	**Micro Average**	**Labels per Document**	**Unlabelled Documents**
Dev Set	0.186	0.305	0.231	9.93	0
Test Set	0.176	0.291	0.219	9.90	0

Table 5.11: Scores for the Neural Network

The early stopping callback stopped the training after 4 epochs. With a patience of 3 in the early stopping callback, this means the performance does not improve anymore after the first epoch.

Inspecting the performance on the individual classes reveals that the F1-score of the neural network on 71 of them is 0, with both precision and recall being 0. On the remaining 10 classes (Table 5.12), the recall is 1 or very close to 1 by assigning almost every document to the class, which results in a low precision but an F1-score > 0, and also leads to an average of almost 10 keywords per document.

Keyword	**Precision**	**Recall**	**Micro Average**	**Occurrence in entire labelled data**
Familie	0.20	1.0	0.34	498
Film	0.06	1.0	0.12	89
Freunde	0.02	1.0	0.044	82
Gefühle	0.48	1.0	0.65	826
Geschlechterrollen	0.13	0.97	0.22	201
Glaube	0.22	1.0	0.36	428
Öffentliche Räume	0.07	1.0	0.13	138
Oper	0.01	1.0	0.02	15
Schreiben	0.33	1.0	0.49	476
Wirtschaft	0.23	1.0	0.38	282

Table 5.12: Results of the Neural Network (F1 > 0)

Thus, the neural network does not work on the data of the Oberfrohna correspondence and will not be further used. Reasons might be the too small number of data points or the ambiguous class labels, or the neural network would have to be built and fine-tuned very differently for this specific use case.

[12] https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping

5.2.5 Stacked Classifier

The stacked classifier combines the three classifiers with the best performance on the development data (Logistic Regression, SVM with sigmoid kernel, and rule-based) and creates the label matrix y based on majority voting.

Naturally, the F1-score on the test data is slightly lower than on the development data, which were used to select the parameters (Figure 5.13).

	Precision	**Recall**	**Micro Average**	**Labels per Document**	**Unlabelled Documents**
Dev Set	0.309	0.441	0.364	8.612	2
Test Set	0.288	0.412	0.339	8.572	3

Table 5.13: Scores for the Stacked Classifier

The proportion of unlabelled documents is relatively low with the stacked classifier. The letters of the test set reported in Table 5.14 remained unlabelled. All three unlabelled letters are characterized by their relatively short length (in comparison, the mean letter length of a letter in the test data is 1,389.21 tokens), indicating that a small number of tokens per document makes the labelling task more difficult for the classifier.

Letter Signature	**Ground-Truth Labels**	**Letter Length**
OBF-420600-001-01	Wirtschaft, Geld	115 tokens
OBF-421023-001-02	Geschenke, Kommunikation, Schreiben	121 tokens
OBF-421024-001-02	Kriegsverlauf, Geschlechterrollen, Hausrat	216 tokens

Table 5.14: Unlabelled Letters

While the micro average is computed across all classes, the F1-score for the individual classes differs vastly. Figure 5.5 visualizes the classifier's performance for each class with an F1-Score > 0 (except for the class "Begegnungen", all F1-scores greater than 0 are better than random).

The classes in Table 5.15 have an F1-score of 0 on the test data. Especially for the class "Freunde", for which a relatively large number of training data is available, it becomes evident that it is hard to distinguish for the classifier and that the training documents have to be reviewed again. The scores in detail for all classes with precision and recall can be found in Appendix B.2.

With the exception of "Freunde", there are very few training documents available for most classes that have an F1-score of 0 on the test data: as there are 782 training documents, each of these keywords was assigned to less than 5.9% of all training documents. For the total of 264 test documents, 24 of the classes with an F1-score of 0 occur in a maximum of 6% of all letters of the test set. The limited number of test documents for these classes does not offer enough information to assess the classifier's overall performance on these classes. Despite the false negatives that the classifier could not assign correctly to their class, it may perform well on other documents belonging to this class. However, this can unfortunately

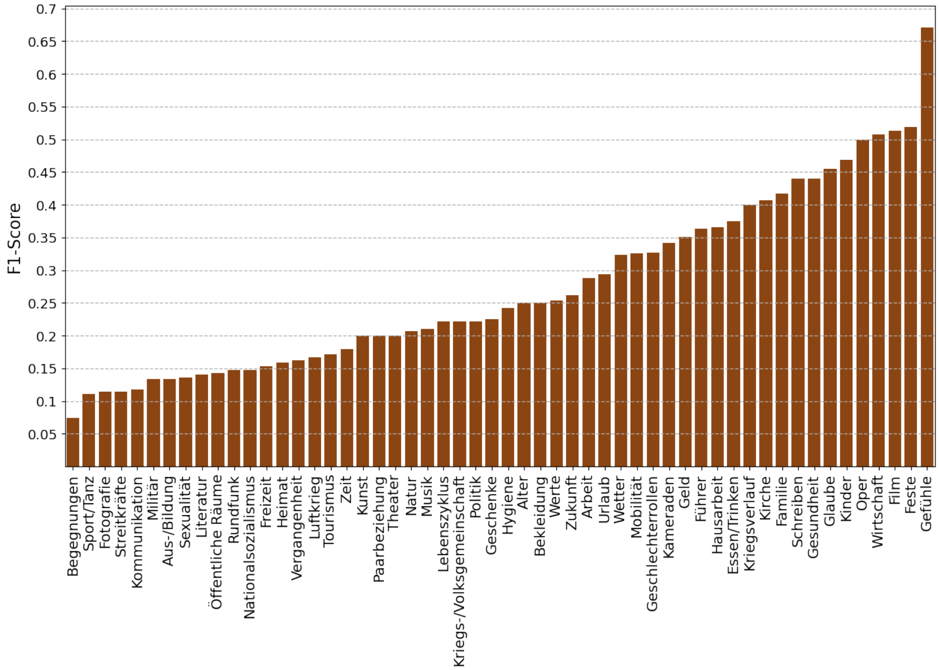


Figure 5.5: F1-Scores per Class of the Stacked Classifier

not be evaluated correctly. Since the stacked classifier outperforms all other models, it will be used to assign keywords to the letters that have not been labelled so far. All classes with an F1-score of 0 on the test data have to be excluded from further analysis.

5.3 Assigning Keywords using the Trained Classifier

The stacked classifier composed of Logistic Regression, SVM, and the rule-based model, applying majority voting, turned out as the most accurate among all the models tested in the experiments. This classifier is applied to all letters that have not been labelled so far. It predicts the keywords of the 514 unknown letters from Hilde (January 1943 to February 1946) and the 808 unknown letters from Roland (February 1943 to February 1946). On average, the stacked classifier assigns 8.12 keywords to each of these documents (6.87 to Roland's and 10.05 to Hilde's). 29 letters remain unlabelled. 26 of these were written by Roland. Figure 5.6 visualizes the distribution of all keywords that the stacked classifier assigned to the 1,322 unlabelled documents.

Since the overall performance that can be achieved is relatively poor (micro-average F1-score of 0.34 on the test data), the obtained labels cannot be used to replace the future manual labelling task of letters when they are published on the AiK website. Using them to answer the humanities-oriented research questions (as

Keyword	Occurrence in Training Data	Occurrence in Test Data	False Positives
Allianzen	2	3	0
Antisemitismus	8	4	0
Ausweise	2	2	0
Baukunst	4	3	0
Behörde	20	3	0
Bekannte	16	13	0
Bevölkerung	4	4	0
Freunde	74	6	0
Gräueltaten	17	1	2
Hausrat	21	16	7
Kriegsfolgen	17	8	0
Kriegsschauplatz	28	14	0
Kultur	25	5	1
Kulturkontakt	33	2	5
Körper	18	5	1
Landschaften	24	2	9
Machthaber	18	13	0
Partei	34	16	1
Praktiken	46	5	4
Private Räume	29	15	6
Rassismus	17	7	0
Reden	9	8	0
Sprache	17	1	1
Status	31	2	0
Waffen	31	3	1
Wissenschaft	33	3	1
Zeitungen	7	4	0

Table 5.15: Classes with F1-Score = 0 by the Stacked Classifier

opposed to the technical questions concerning the highest achievable F1-score on classification) posed at the outset of the thesis is also problematic, and extreme caution must be exercised. The low precision of 0.29 indicates a high number of false positives/incorrectly assigned keywords. If trends in the frequency of a keyword can be observed over time, they are not necessarily reliable. Instead, it must be checked manually whether Hilde and Roland were indeed writing more about a certain topic in a certain period. Also, one must keep in mind that the F1-score of 0.34 was measured on the test data, while the test data is not necessarily representative of the unlabelled letters. Hilde and Roland wrote these letters during a later time frame, commencing from January/February 1943. It is plausible that the topics of emphasis could have evolved in contrast to the period when they wrote the letters which are part of the test set (from May to December 1942). Additionally, their writing styles and the utilization of specific words could have changed. Moreover, only the letters until August 1943 were proofread by the

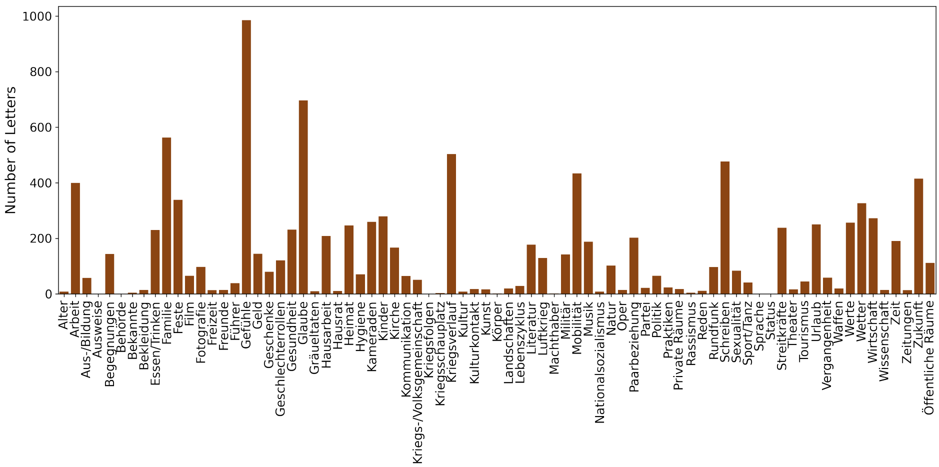


Figure 5.6: Distribution of Classes assigned to Unlabelled Data

time of the pre-processing. Starting from September 1943, there are still HTR errors in the transcripts.

In addition to the keywords assigned by the classifier that must be treated with caution, the already present keywords are not necessarily reliable due to how they were created. While the already present keywords necessarily had to be treated as ground-truth for training the classifier even though they are not necessarily accurate, conducting a content analysis now requires even more caution.

Furthermore, it has to be kept in mind that due to the “incomplete” months in the correspondence (see Figure 4.1), trends that could potentially be investigated might remain uncovered. On the one hand, a certain keyword could under different circumstances be assigned to many letters which in fact could be lost. On the other hand, a certain topic might have just been very relevant between the two spouses when Roland was on furlough or when letters were not delivered by post. Letters from this period could reflect the importance of the topic, but due to personal contact between Roland and Hilde or lost letters there is no written evidence from this period.

5.4 Interpretation of the Keywords

5.4.1 Differences between Hilde and Roland

The first research question is dedicated to differences (or also similarities) in the writing habits and topic focuses of Hilde and Roland that can be deduced from the keywords. Some observations could already have been made in the exploratory data analysis in Chapter 3.6. Hilde on average wrote slightly longer letters than Roland (1,365 vs. 1,041 words), but Roland likely wrote slightly more letters than Hilde (1,502 vs. 1,129 if we assume that the loss of Hilde’s letters is not severely greater than the loss of Roland's). This assumption is supported by the fact that the total word count for both of them is very similar. The higher average number of keywords per letter (in the entire correspondence on average 7.16 for Roland's letters and 9.10 for Hilde's letters; median 7.0 keywords for Roland and 9.0 keywords for Hilde) indicates that Hilde's letters are longer and cover a wider range of topics, while Roland's individual letters are shorter and cover fewer topics, and he is more likely to start a new letter to address a new topic.

As for the most frequent keywords for each of the two spouses, there are some overlaps. Table 5.16 summarizes the five most frequent keywords for both Roland and Hilde, giving information about how many letters were labelled with a certain keyword.

	Roland	**Hilde**
Total letters	1502	1129
Gefühle ("feelings")	971	841
Familie ("family")	382 (rank 7)	679
Glaube ("belief")	638	487
Schreiben ("writing")	466	487
Arbeit ("work")	466	240 (rank 14)
Mobilität ("mobility")	425	328 (rank 10)
Wetter ("weather")	241 (rank 11)	398

Table 5.16: Most frequent Keywords per Author

"Gefühle" ("feelings") is the most frequent topic for both spouses, which is only logical because the letters are first and foremost love letters. However, the term "feelings" must not be understood too narrowly in the AiK context. In the project documentation (Fahnenbruck et al. 2023) it is defined as „Umfasst den Austausch von emotionalen Erfahrungen, nicht aber die Bewertung von Ereignissen oder die Thematisierung von Normen (siehe Werte)" ("Includes the exchange of emotional experiences, but not the evaluation of events or the thematization of norms (see values)"). It naturally includes especially feelings of love and affection, but also a spectrum of other feelings articulated within the letters. These emotions can be directed towards the recipient, as well as other individuals or events referenced in the correspondence. In the top five of Roland's and Hilde's most frequent keywords, also "Schreiben" ("writing") and "Glaube" ("belief") overlap. Both are themes that strongly connect Hilde and Roland over the distance. The writing of the letters itself is a recurring theme, and both Hilde and Roland sought hope in their faith in God for a soon reunion. Many letters of especially Hilde contain prayers, which are usually closed with an "Amen.". There are 230 letters of Hilde which contain the string " Amen." (with whitespace and period), indicating that at least 20.37% of her letters contain prayers. (In Roland's letters, the string " Amen." occurs only 11 times.) The general topic "Glaube" ("faith") can be found in 638 letters of Roland and 487 letters of Hilde, as far as the keywords can be trusted.

The words in the top 5 that do not overlap between the two spouses are "Arbeit" ("work") and "Mobilität" ("mobility") for Roland and "Familie" ("family") and "Wetter" ("weather") for Hilde. Their high frequency can be justified by the everyday lives of Hilde and Roland. One of the main motifs of field post letters was to maintain relationships and let each other participate in daily life (Humburg 2011, 78). Work (ranked 14$^{\text{th}}$ for Hilde) was the most important part of Roland's life during deployment, whereas Hilde kept living with her parents during Roland's absence, making her family (ranked 7$^{\text{th}}$ for Roland) the closest caregivers present, with a correspondingly high placed value in her life. Weather (ranked 11$^{\text{th}}$ for Roland) is according to Humburg (2011, 79) one of the most common standard topics in field post letters, trying to bridge some of the spatial distance. The disparity in mobility between Roland and Hilde, defined as "Umfasst alle Fortbewegungsweisen, Verkehrsmittel und Ortswechsel, wie Züge, Zugfahrt, Truppenbewegung, Truppenverlegung" ("Includes all modes of locomotion, transportation, and change of location, such as trains, train travel, troop movement, troop relocation") (Fahnenbruck et al. 2023) and ranked 10$^{\text{th}}$ for Hilde, can be attributed to their differing geographical constraints. Hilde's mobility was primarily limited to her hometown of Oberfrohna, with occasional visits to her in-laws in Kamenz, located 130 kilometers away. While she relied on Roland's letters to stay updated on his location changes, his war service led him to various locations. In the letters he often described his travels, for example, when he was on his way back from furlough in 1942:

Nach dem nun schon gewohnten Gesuche u. Gepacke sitze ich im Wartezimmer in Türnähe, damit ich unter denen bin, die mitkommen. [...] Die Fahrt ging gut bis hierhin. Ganz pünktlich setzte mein Zug sich in Bewegung. Er stand schon da ganz schwach besetzt, ich hatte ein Polsterabteil ganz für mich fast bis Dresden. [...] Wir bekamen einen Sitzplatz im überfüllten Zuge – mußten in Prag aber wieder herauskrabbeln, hatten wieder Glück und fuhren gepolstert bis Wien [...].

After the now already usual searching and packing I am sitting in the waiting room near the door, so that I am among those who come along. [...] The trip went well up to here. My train started moving quite on time. It was already very lightly occupied, I had an upholstered compartment all to myself almost to Dresden. [...] We got a seat in the overcrowded train—but had to crawl out again in Prague, were lucky again and rode upholstered to Vienna [...].
(Fahnenbruck et al. 2023, OBF-421014-001-01)

(The exact counts of keywords across all letters in already labelled and unlabelled letters as well as all keyword definitions can be found Appendices A.2 and A.3.)

5.4.2 Consistently high Keyword Frequencies

According to Humburg (2011, 79), there are some standard topics in all field post letters between spouses or family members. These usually remain stable throughout the entire correspondence. The standard topics Humburg (2011, 79) mentions are also very common in Hilde and Roland's letters. Standard topics include writing and receiving letters (captured by the keyword "Schreiben"), expressing love (keyword

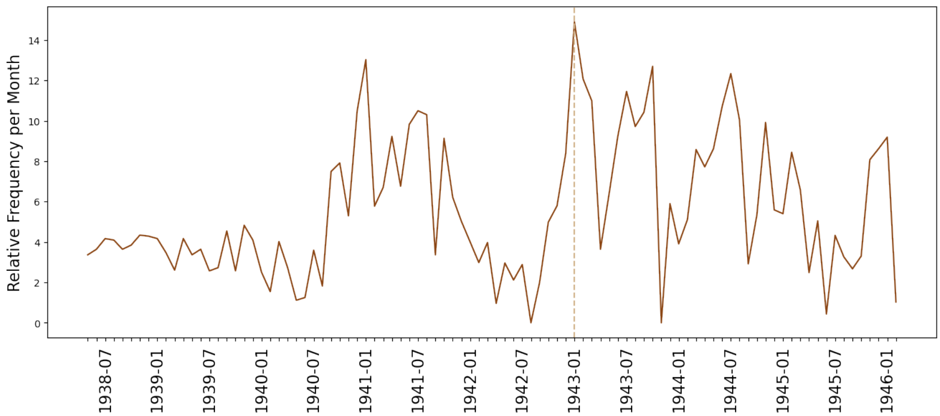


Figure 5.7: Relative Frequency of “Gefühle”

“Gefühle”), attempts to reassure the spouse, military service (keyword “Militär”), weather (keyword “Wetter”), and the daily life at home, along with the issues concerning the family and the upbringing of the children. Hilde did write a lot about her daily life in her home village (keyword “Hausarbeit”) and her family (keyword “Familie”). However, the upbringing of children was not yet a matter of discussion.

Due to missing letters and incomplete months, there are no keywords whose frequency is consistently high and there are large fluctuations in the line chart for each keyword. For instance, even the line chart for “Gefühle” (“feelings”), the most common keyword in the entire correspondence and perhaps the most important topic for Hilde and Roland, fluctuates. However, it cannot be assumed that feelings played a much more important role in, for instance, 1943 than in 1942 (see Figure 5.7[13]).

To gain a more comprehensive understanding of which keyword frequencies exhibit less variability than others, beyond solely relying on visual comparisons of their line charts, the standard deviations of the normalized keyword frequencies across all months of the correspondence can be compared. To make the relative frequencies of the different keywords comparable, they have to be scaled first (depending on the total frequency of a keyword, also the relative frequencies per month vary, as can be seen on the y-axes of the different line charts). Therefore, a min-max transformation (scaling all relative frequencies between 0 and 1) is applied. Using the scaled relative frequencies, the standard deviations could be calculated. However, to detect consistently high keyword frequencies, only standard deviations of keywords with high mean frequencies are interesting. Rare keywords would have a low standard deviation and a low mean as their frequencies are consistently close to 0. Common keywords, however, can be expected to have a low standard deviation and a high mean as their frequencies are consistently close to 1.

[13]The dashed lines in January 1943 indicate the transition from the semi-manually assigned labels to the predicted labels.

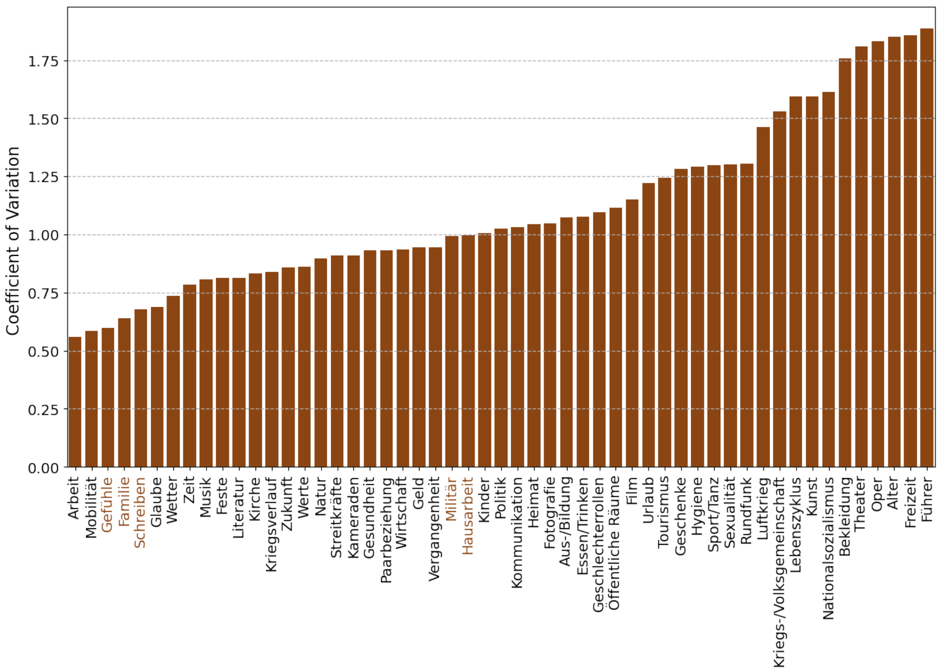


Figure 5.8: Coefficients of Variation

To find the keywords with low standard deviations and high means of relative keyword frequencies, the coefficient of variation (CV) for all keywords with an F1-score > random is calculated. The CV indicates the ratio of the standard deviation to the mean. Keywords with a high mean frequency and low standard deviation therefore will have the lowest coefficients of variation, indicating their frequencies are consistently high. The coefficient of variation is the quotient of standard deviation and mean:

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

where:

σ = standard deviation

μ = mean

The CVs of the normalized relative frequencies per month for the keywords with an F1-score > random on the test data can be seen in Figure 5.8. The graph shows that the most frequent keywords (see Chapter 5.4.1) are also the most consistent ones throughout the entire correspondence. As for the keywords which can according to Humburg (2011) be expected to be consistent throughout the correspondence, “Gefühle”, “Familie”, and “Schreiben” in fact have a very low ratio of mean and standard deviation. "Militär" and "Haushalt" (in the middle third of all CVs) are less consistent than expected. The rightmost keywords, on the other hand, occur only in specific months, as it will demonstrated below with "Führer".

The CVs must be viewed with caution again. A high keyword consistency, as represented by the CV, is closely tied to the relevance of the corresponding topic for Hilde and Roland. The keywords that are consistent throughout the entire correspondence according to the CV are also better represented in the training data and therefore the model can make more consistent predictions for them. More consistent predictions result in a higher CV again.

Thus, from Distant Reading alone one can certainly assume that there were "standard topics" in the correspondence. (This becomes even more evident upon directly reading the letters of the correspondence.) Because of the high frequency of the corresponding keywords of these standard topics, it is also relatively easier for the classifier to predict them consistently.

5.4.3 Trends of Topics

Apart from the keywords which are part of the hypotheses and will be investigated below in detail, there might be some other keywords which reveal trends over time. The keywords "Essen/Trinken", "Kameraden", and "Führer" are examples of keywords which appear to have particularly interesting progress curves ("not interesting" in this case would mean, for example, a constant fluctuation over the whole period of correspondence, where a keyword is alternately more or less interesting, but no trend can be observed[14]).

Essen/Trinken The topic "Essen/Trinken" ("eating/drinking"; Figure 5.9) only gains relevance over time. One reason for the almost absent presence of the keyword until 1940 in the original labelled data may of course be that it was less in the focus of the annotators. Disregarding this problem in the labelling process and assuming that the curve reflects the actual importance of the topic for Roland and Hilde, this would mean that eating and drinking are not discussed in the letters until the relationship turns into a long-distance one. Eating and drinking may have been practised together before or may have been too "mundane" to discuss in the letters. Later, it turns into one of the "standard topics" that couples typically share during a deployment to maintain normalcy and keep each other involved in their daily lives (Humburg 2011, 79). The topic of eating and drinking is further investigated later (Chapter 5.4.4 on the keyword "Wirtschaft").

Kameraden Something similar can be observed with the keyword "Kameraden" ("comrades"; Figure 5.10), which is, according to the AiK definition, dedicated

[14]At this point it should be emphasized again that topics with "uninteresting" curves can of course also have peaks or be more or less present in Hilde and Roland's letters in a certain, conspicuous period. The fact that these keywords are not discussed further here only means that nothing particularly conspicuous can be observed in the visualization of the relative keyword frequencies based on existing and predicted keywords.

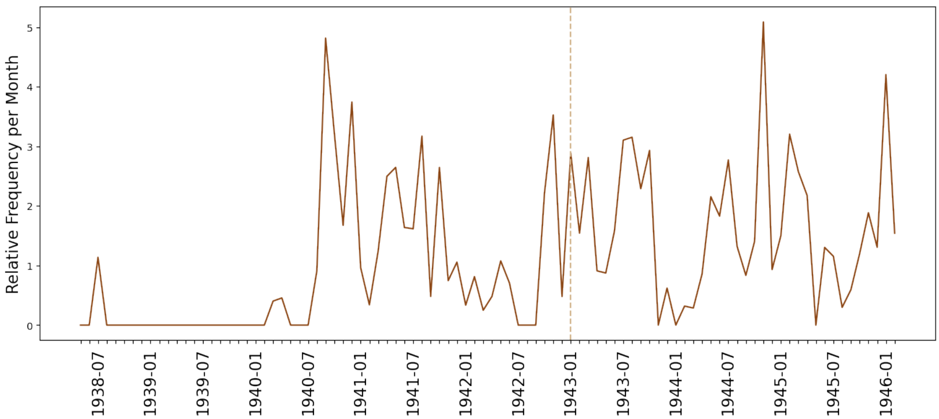


Figure 5.9: Relative Frequency of "Essen/Trinken"

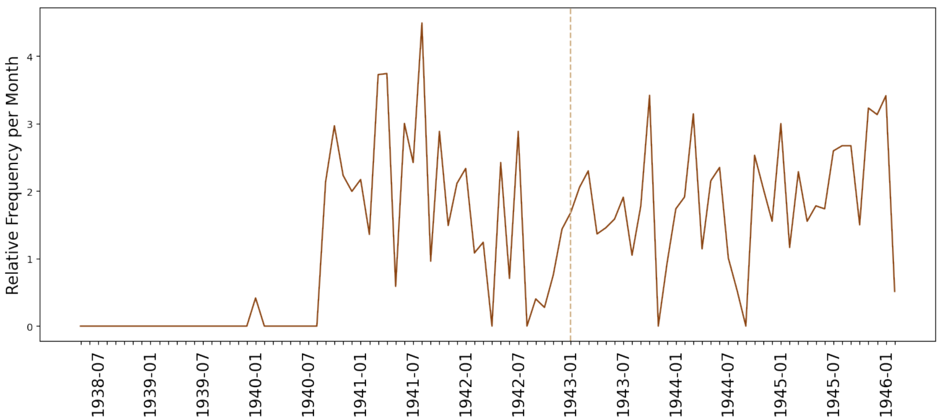


Figure 5.10: Relative Frequency of "Kameraden"

to „Kameraden und Kameradinnen im Militär und in den zivilgesellschaftlichen Organisationen der NSDAP, Kameradschaft, Kameradschaftsehe, Kameradschaftsgefühle, Appelle an Kameradschaftlichkeit" ("Comrades in the military and the civilian organizations of the NSDAP, comradeship, comradely marriage, feelings of comradeship, appeals to comradeship"). Only with Roland's deployment in 1940, the topic became relevant. Before moving to military barracks, Roland simply did not have any comrades to write about. The higher importance of comrades for Roland than for Hilde is also reflected in the number of letters labelled with this keyword for each of them (339 for Roland vs. 100 for Hilde).

Führer The graph for the keyword "Führer" (Figure 5.11) has an outstanding peak in January 1943 (in this month 11 out of 64 letters are labelled with the keyword). Having a closer look in the labelled letters reveals that most of them

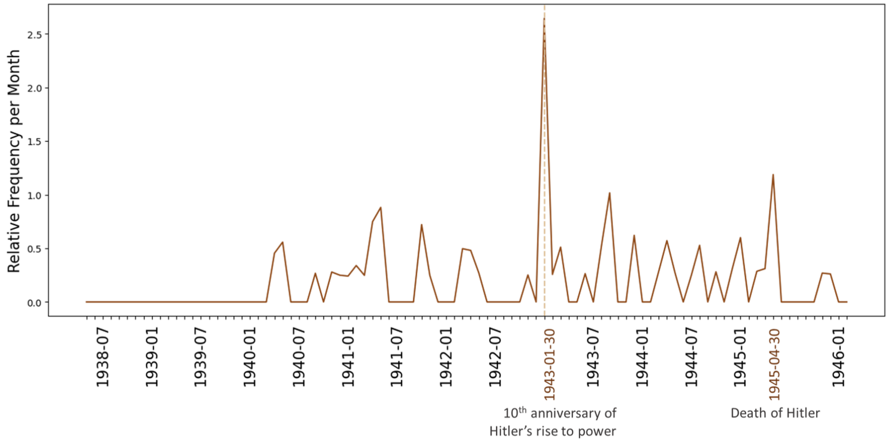


Figure 5.11: Relative Frequency of "Führer"

address Hitler's 10-year takeover. For instance, both Hilde and Roland talked about preparations for the anniversary, Roland not at all uncritically:

Für morgen muß ich auch noch einiges zum 10 jährigem Gedenken der nationalen Erhebung zusammenstellen, 30. Januar, es soll wohl ein Staatsfeiertag sein hier. Bin gespannt, wenn der Führer spricht, was er nun zu sagen hat.

For tomorrow, I also have to put together something for the 10-year commemoration of the national uprising, January 30, it should probably be a national holiday here. I am curious when the Führer speaks, what he has to say now.
(Fahnenbruck et al. 2023, OBF-430126-002-01, unpublished)

Heute kamen in unsre Geschäftsstelle auch Anweisungen über die Ausgestaltung des 30. Januar, des 10. Jahrestages der Machtübernahme. Am Sonnabend soll Arbeitsruhe sein – am Freitag schon Kundgebungen – drei Tage soll gefeiert werden. Die Propagandamaschine läuft wieder auf Touren – und das ist ja auch nötig, um jedermann diesen Tag als einen Wendepunkt zu einem besseren Dasein erscheinen zu lassen.

Today our office also received instructions on the organization of January 30, the 10th anniversary of the seizure of power. On Saturday there is to be a rest from work—on Friday already rallies—three days are to be celebrated. The propaganda machine is running at full speed again—and that is also necessary to make this day appear to everyone as a turning point to a better existence.
(Fahnenbruck et al. 2023, OBF-430125-001-02, unpublished)

There is also a small local peak in May and June 1945, after Hitler's death, but this does not stand out much more than other local peaks. When looking more closely at Roland's letters (unfortunately there are no letters from Hilde from May 1945), it also becomes clearer why it can be difficult for the classifier to detect the class "Führer". The letter from May 2, 1945, which Roland wrote after learning from Hitler's death, was labelled with the keywords "Familie" ("family"), "Gefühle" ("feelings"), "Gesundheit" ("health"), "Glaube" ("belief"), "Kirche" ("church"), "Mobilität" ("mobility"), and "Zeit" ("time), therefore false negative for "Führer", but also "Politik" ("politics"), "Kriegsverlauf" ("course of war") and "Machthaber" ("rulers"). Roland used metaphors in his bewilderment and turned to God. He tried to frame what happened by attributing it to destiny, but he hardly talked openly about it, which makes it hard for the classifier to correctly predict the keywords:

[...] und wenige Stunden später war sie uns bekannt, die Tatsache, die für das kommende Geschehen, für unser Volk die einschneidendste der möglichen Tatsachen ist. [...] Wir stehen still und stumm an der Bahre dieses Mannes – dieses furchtbaren Mannes – ein Werkzeug Gottes auch er, bestimmt, ein furchtbares Gericht über unser Volk zu bringen – ein Finger Gottes.

[...] and a few hours later it was known to us, the fact which is the most drastic of the possible facts for the coming events, for our people. [...] We stand still and silent at the bier of this man—this terrible man—an instrument of God he too, destined to bring a terrible judgment upon our people—a finger of God.
(Fahnenbruck et al. 2023, OBF-450502-001-01, unpublished)

5.4.4 Examining the Hypotheses

H1: The frequency of the keyword "Luftkrieg" ("air war") is directly tied to real-time events, showing an increase whenever actual attacks occur Hypothesis 1 says that the graph for the keyword "Luftkrieg" ("air war") reflects actual happenings of World War II at certain points of time. Out of the keywords related to war (also "Kriegsschauplatz" ("battlefield"), "Kriegsverlauf" ("course of war"), "Kriegsfolgen" ("consequences of war")), „Luftkrieg" appears to be the most promising to display real war events in its graph. It might reflect real-time events, while "Kriegsverlauf" and "Kriegsfolgen" describe the war from a long-term perspective. Also, in comparison to "Kriegsschauplatz", it aims at temporal events rather than locations, which might be better visible in a graph that visualizes a timeline. If the hypothesis can be verified, it demonstrates that Distant Reading of the correspondence allows for establishing connections with real-time events and enables targeted searches for letters dedicated to the topic "Luftkrieg". In the stacked classifier, the topic achieves an F1-score of 0.16 (randomly assigning the keyword would result in an F1-score of 0.05). Figure 5.12 depicts the keyword's relative frequency over time. The two events related to the peaks in the second half of 1940 and the beginning of 1945 within the context of the correspondence are the Battle of Britain, one of the Luftwaffe's (Nazi Germany's air force) largest

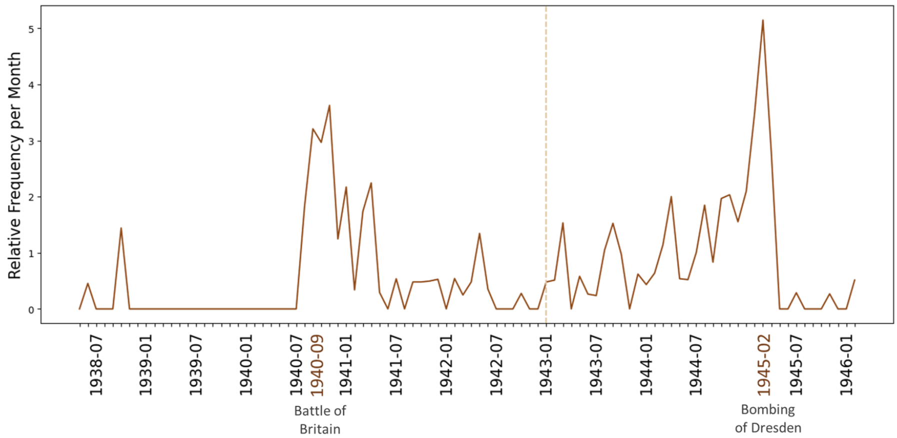


Figure 5.12: Relative Frequency of "Luftkrieg"

attacks, and the bombing of Dresden and Chemnitz in Saxony, which was Hilde and Roland's home state.

In November 1940, 13 out of 36 letters are labelled with the keywords "Luftkrieg". Roland, at this time situated in Barkelsby, told his wife of a typical night in military service during the British counteroffensive:

Das bedeutet: daß unser Dienst, der Dienst der Soldaten hier zumal, erst nachts recht beginnt, wenn der Tommy einfliegt. Eine kleine Kostprobe letzte Nacht: 4 mal aufgestaenden [sic] wegen Alarm — die Soldaten müssen dann an die Geschütze, ich muß dann in die Schreibstube zur „Alarmwache." Nur in einem Falle überflog uns ein Flugzeug, das auch beschossen wurde.

That means: that our service, the service of the soldiers here in particular, only really begins at night, when the Tommy flies in. A small sample last night: 4 times we got up because of an alarm—the soldiers then have to go to the guns, I then have to go to the writing room for "alarm watch." Only in one case an airplane flew over us, which was also shot at.
(Fahnenbruck et al. 2023, OBF-401102-001-01)

("Tommy" here is emblematic of the aircraft of the RAF. In World War I, the United States used the Thomas-Morse S-4 Scout biplane aircraft, abbreviated as "Tommy". After the end of World War I, however, these were no longer used in war (Dwyer 2013).)

The "Battle of Britain" began on August 13, 1940, when the Nazis wanted to gain air supremacy over Great Britain. While Nazi Germany propagated the complete superiority of the German Air Force, shortcomings in equipment and in training of the German Luftwaffe quickly became apparent. On September 15, 1940, the climax of the battle of Britain, the German Luftwaffe lost 56 aircraft, but Nazi German continued to attack large British cities such as London (Scriba 2015). Until the attack on London, the RAF's counteroffensive had been directed primarily at military targets. After that, the British population began to call for an end of "passive martyrdom" and for war against all of Germany, including civilians (Süß 2011, 106). Hilde and Roland now address these counterattacks in their letters, and they are also reflected in the progression curve of the keyword.

The Battle of Britain ended in the spring of 1941 when Hitler needed all aircraft for the attack on the Soviet Union. All in all, in the battle around 2,000 German pilots died. 2,600 more were either missing or in British captivity. Moreover, Nazi Germany lost 2,256 aircraft, and another 867 were damaged by more than 10% (Scriba 2015).

In February 1945, 12 out of 31, and in March 1945, 18 out of 33 letters are labelled with "Luftkrieg". In the total of 30 letters labelled with the keyword "Luftkrieg" in February and March 1945, there would be a large number of interesting passages that could be given here as examples. To keep it short, here is the beginning of Hilde's letter from February 14:

Und es wurde fürchterlich! Tausende mögen doch gekommen sein, es bebte alles, krachte und dröhnte. Ich habe gezittert vor Angst und konnte nur dasitzen erstarrt die Hände gefaltet. Es ist schrecklich so auf den Tod zu warten, wenn man nicht geschützt ist vor dem Unheil von oben! Was ist ein Keller bei Volltreffer. Über 2 Stunden hausten sie in der Umgebung. Chemnitz-Altstadt und Kappel voll brennen lichterloh, das sah ich als ½ 12 Uhr Entwarnung kam vom Fenster aus blutroter Himmel! Das Licht war weg.

And it became terrible! Thousands may have come, it all shook, cracked and roared. I was trembling with fear and could only sit there frozen hands folded. It is terrible to wait for death when you are not protected from the disaster from above! What is a basement at direct hit. For more than 2 hours they dwelled in the area. Chemnitz-Altstadt and Kappel completely burn ablaze, this is what I saw when all-clear came at half past 11 o'clock from the window blood-red sky! The light was gone.
(Fahnenbruck et al. 2023, OBF-450214-002-01, unpublished)

In this passage, the partial lack of punctuation is very noticeable. Commas are sometimes missing in Hilde's sentences, but here some sentences become ungrammatical, which is usually not the case in other letters. Her writing reflects how Hilde wrote the letter in greatest tension when she hastily told Roland how she had just feared for her life.

The attack on Dresden in the night of February 14, 1945, which Hilde describes here, was one of the most devastating air raids in World War II. The Allies deliberately bombed German civilians to break their morale. What Hilde perceived as thousands of aircraft were indeed 773 British bombers. In this night, 80,000 apartments in Dresden were destroyed. Until February 15, Dresden was bombed again twice by the US Air Force. In total, up to 25,000 civilians died (Scriba 2022).

The peak in February/March 1945, during the air raids on Saxony, on the one hand, confirms that the classifier is indeed able to correctly label letters whose transcripts have not been corrected yet. Out of the 30 relevant letters in February

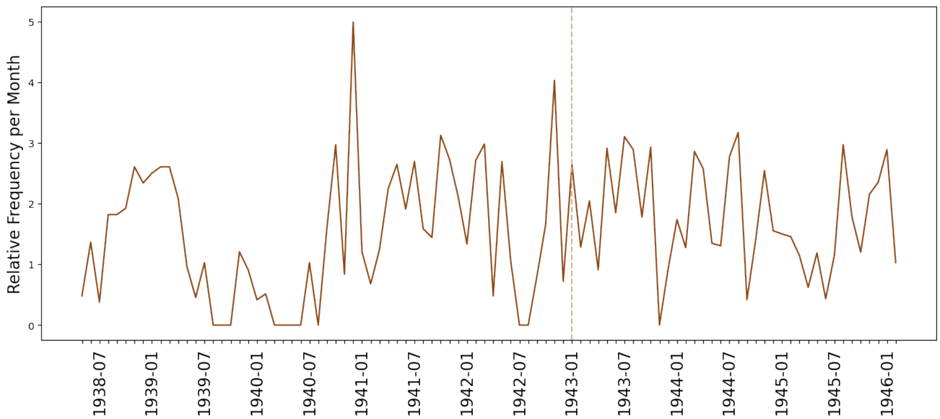


Figure 5.13: Relative Frequency of Keywords related to Culture

and March 1945, two-thirds (20) are written by Hilde, even though the CER of Transkribus is slightly higher for her handwriting than for Roland's.

On the other hand, what the line chart cannot prove is that the classifier can fully capture the topic "Luftkrieg". The severity of the attacks on Saxony does explain the extreme high rise, however, what the graph cannot show is whether the two time periods with the highest relative keyword frequencies were indeed those when the topic of air war was most present for the spouses. False negative letters for the keyword "Luftkrieg" or gaps in the correspondence could lead to the fact that further peaks are overlooked.

H2: The frequency of keywords concerning culture (Kultur ("culture"), Kunst ("art"), Theater ("theatre"), Oper ("opera"), Musik ("music")) decreases over the course of the war. Roland was a musician. The letters of the first half of the war, which have already been processed by AiK members, reveal that music and culture/art played an important role in Hilde and Roland's life. Towards the end of the war, the economic situation and the living conditions got worse and worse. This leads to the assumption that arts then played a subordinate role and that plain survival was in the foreground. The hypothesis should provide information on whether this is also reflected in the keywords related to culture.

From the five keywords related to culture, unfortunately, the most obvious one, "Kultur" ("culture") itself cannot be further considered as the classifier achieves an F1-score of 0 on the test data. A combination of the remaining keywords "Kunst", "Theater", "Oper", and "Musik" can be seen in Figure 5.13. Overall, there is no clear trend visible. The peak in December 1940 was solely caused by a high number of letters labelled with "Musik". Reading these 18 out of the 55 letters from December 1940 proves that the keyword was correctly assigned to them. However, the topic of music is dealt with in a wide variety of contexts, and it seems rather like a coincidence that so many letters are labelled with it in this month. The letters are dedicated to Hilde's "Singstunde" (choir rehearsal) (Fahnenbruck et al. 2023, OBF-401205-001-01, OBF-401205-002-01), a Christmas party Roland was asked to play music at (Fahnenbruck et al. 2023, OBF-401209-001-01, OBF-401216-002-01), childhood memories of a school concert (Fahnenbruck et al. 2023, OBF-401220-001-01) and in general, due to the season, Christmas music (Fahnenbruck et al. 2023, OBF-401225-001-01). On December 21, 1940, Hilde and her parents got their first radio. Now Hilde and Roland could establish listening to the radio together as a shared practice in their long-distance relationship (see Bergerson, Fahnenbruck, and Hartig 2019, 270):

Aber wir werden die Feiertage Musik haben — Musik!!! Herzlieb! Du!! Ach — kannst Du meine große [Fre]ude darüber verstehen? Du!! Was Du hören magst, was Du hören wirst, das soll ich auch hören?!

But we will have music over the holidays—music!!! Dear heart! You!! Oh—can you understand my great joy about it? You!! What you like to hear, what you will hear, I should hear too?!
(Fahnenbruck et al. 2023, OBF-401221-002-01)

From then on, they regularly talked about the music both listened to on the radio (Fahnenbruck et al. 2023, OBF-401226-001-01, OBF-401226-002-01).

Due to its big increase in December 1940 caused by a combination of several factors, the topic "music" seems to have more relevance at this time for Hilde and Roland than, for instance, in December of the following years. However, nothing can be deduced from this about the big picture and a declining priority of culture (or even just music) throughout the course of the war. H2 has to be rejected.

H3: The frequency of the keyword "Wirtschaft" ("economy") increases towards the end of the war, as more economic and financial concerns are discussed. In contrast to the previously studied culture-related keywords, the assumption for the keyword "Wirtschaft" is that its frequency rises towards the end of the war. When the economic situation in Germany worsened, it might have become a more relevant topic for Hilde and Roland, especially concerning sourcing food, as food became increasingly scarce.

The keyword "Wirtschaft" can be predicted with an F1-score of 0.41 on the test data. In comparison to most other keywords, this is relatively high (randomly assigning letters to this class would result in an F1-score of 0.19), thereby enhancing the reliability of relative keyword frequencies per month. The hypothesis suggests that the economic worries, which become greater as the war progresses, are also reflected in the letters. The corresponding plot (Figure 5.14), however, cannot confirm this. One can observe in Figure 5.14 that economy was a less relevant topic before Roland's deployment but became more frequently discussed afterwards. One possible explanation is that the topic itself does not necessarily indicate economic worries. It is possible that Hilde and Roland did worry more about their financial and economic situation by the end of the war, however, the keyword itself does not differentiate between worries and positive or neutral statements. Apparently, the topic was most present at the beginning of 1942, as in February and March

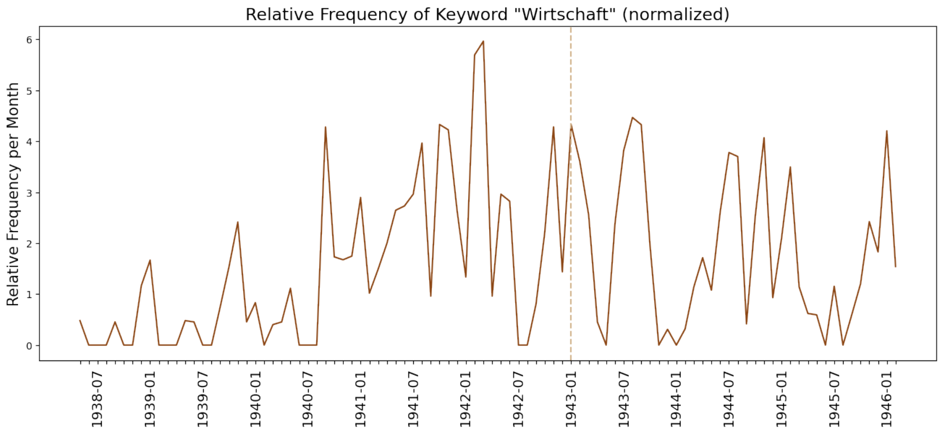


Figure 5.14: Relative Frequency of "Wirtschaft"

1942, 45 out of 86 letters were labelled with "Wirtschaft". A closer look at the letters of these two months, however, reveals very little evidence that the topic was particularly important then. In addition to a large number of letters falsely assigned to the class "Wirtschaft", there are some letters that do address financial matters and ration cards, as well as other topics related to the economy, such as economic restructuring in favour of the arms industry:

Meine erste Umschau und Nachfrage galt einem Paar Stiefelein für mein Herzlieb. Gibt es gar nicht mehr, nur auf Bezugschein, dann kosten sie 80 Mark. Aber diesen Bezugschein erhalten wir nicht: Und ohne ihn kosten die Stiefelein das Doppelte, weil das Leder im Preis rapid gestiegen ist.

My first search and inquiry was for a pair of boots for my Sweetheart. They are no longer available, only on a ration coupon, then they cost 80 marks. But we don't get this ration coupon: and without it, the boots cost twice as much, because the price of the leather has risen rapidly.
(Fahnenbruck et al. 2023, OBF-420220-001-01)

Die ledigen Frauen müssen alle in die Rüstung. Etliche müssen schon am 1. April anfangen bei H.'s, Julius K. und Ernst H. Die Stimmung ist denkbar bedrückt unter allen. Ich bin gespannt, was aus Mutter wird. Ihr Betrieb wird in 4 Wochen schließen, ungefähr. Dora P. ist, [sic] bei H., wird auch geschlossen. Von Hilde K., meiner einstigen Kollegin, erfuhr ich, daß W.s auch schließen werden am 1.April. Nicht 20 Betriebe, sondern 70 sind es im Kreis Chemnitz.

The unmarried women all have to go to the arms industry. Some of them have to start on April 1 at H.'s, Julius K., and Ernst H. The mood is very depressed among all of them. I am curious to see what will become of Mother. Her business will close in 4 weeks,

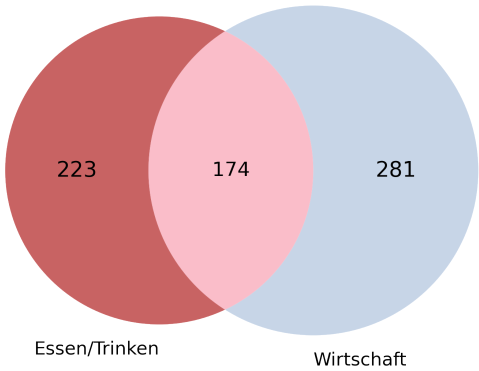


Figure 5.15: Overlap of the Keywords “Essen/Trinken” and “Wirtschaft” on the entire dataset

approximately. Dora P. is at H., will also close. From Hilde K., my former colleague, I learned that W.’s will also close on April 1. It is not 20, but 70 businesses in the district of Chemnitz.
(Fahnenbruck et al. 2023, OBF-420327-002-01)

What is particularly striking when reading the letters of these two months labelled with “Wirtschaft” is how intertwined economy and eating/drinking seem to be. A large number of the letters from February and March 1942 deal with food ration cards and a shortage of food. In the entire correspondence (letters labelled by both project members and the text classifier), almost 44% of all letters assigned to the class “Essen/Trinken” (“eating/drinking”) also have the label “Wirtschaft”, and 38% of all letters assigned to “Wirtschaft” also have the label “Essen/Trinken” (Figure 5.15). While the overview over the entire corpus suggests that there is typically a large overlap between these two topics, a similar pattern is not evident at all for February and March 1942, as only four letters are labelled with “Essen/Trinken” (Figure 5.16). February and March 1942 were not labelled by the classifier yet but the project members and then by summarizing the old keywords to concept terms. A look at each individual letter of these two months would be required now to double-check whether the keywords “Wirtschaft” and “Essen/Trinken” were assigned correctly in the process of creating new keywords as concept terms. However, this would then not be limited to these two months, where a problem has now clearly been identified, nor to these two keywords. Instead, one is again faced with the problem that the entire labelled documents would have to be rechecked to ensure ground-truth. Since this is out of the scope of the thesis, two example letters from Hilde should be used to examine why they, like many others, were not assigned to the “eating/drinking” class during the (semi-) manual labelling process:

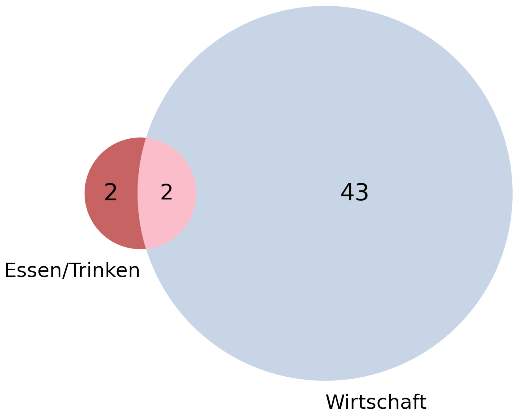


Figure 5.16: Overlap of the Keywords “Essen/Trinken” and “Wirtschaft” in February/March 1942

Und die paar Marken reichen doch weder hin noch her. Wenn Du von den Fleischmarken zum Essen gehst, dann hast Du nichts zum Brot. Kaufst Du Wurst, dann kannst Du nicht Mittagessen.

And the few cards are neither enough back nor forth. If you get lunch from the meat ration cards, then you have nothing for supper. If you buy sausage, then you can't have lunch.
(Fahnenbruck et al. 2023, OBF-420216-002-01)

Feldküchengericht gibt's am Donnerstag. Schnittbohnen mit Fleischklößchen und Sardellensoße. Ein ganz nettes Gericht. 50 g Fleischmarken, 10 g Fett, 40 g Brot. So wird man seine Marken los ––– fühlt aber im Magen noch eine sehr verdächtige Leere.

Field kitchen meal we have on Thursday. String beans with meatballs and anchovy sauce. Quite a nice dish. 50 grams meat ration cards, 10 grams fat, 40 grams bread. That's how you get rid of your cards ––– but you still feel a very suspicious emptiness in your stomach.
(Fahnenbruck et al. 2023, OBF-420319-002-01)

The previously mentioned ration cards appear to be the main reason why many letters dealing with both economy and eating/drinking are only assigned to “Wirtschaft”. The two example letters were originally labelled with the keywords in Table 5.17.

The original keywords of the two example letters which were later combined under the concept term “Wirtschaft” are “Kriegswirtschaft” (“war economy”) for the letter OBF-420216-002-01 and “Einkauf” (“purchase”) for OBF-420319-002-01. The first letter, OBF-420216-002-01, might have been assigned only to the class

OBF-420216-002-01:	Ariernachweis (Aryan certificate)	→	Ausweise (identifications)
	Gemeinde	→	[deleted]
	Gesundheit (health)	→	Gesundheit (health)
	Kriegswirtschaft (war economy)	→	**Wirtschaft** (economy)
	Paket (parcel)	→	Schreiben (writing)
	Urkunden (certificates)	→	Ausweise (identifications)
	Vorwürfe (accusations)	→	Werte (values)
OBF-420319-002-01:	**Einkauf** (purchase)	→	**Wirtschaft** (economy)
	Kinderbetreuung (child care)	→	Kinder (children)
	Reise (travel)	→	Mobilität (mobility)

Table 5.17: Old and New Keywords of Example Letters

"Wirtschaft" because the annotator only focused on the topic of war economy, from which the worries about the procurement of food first resulted. The annotator did not consider food itself relevant in this context. Given that the keywords "Essen/Trinken" and "Wirtschaft" are very often assigned together, as well as Hilde's description of her worries about not having enough to eat for either lunch or supper, the assignment of the keyword "Essen/Trinken" seems appropriate. However, this again is only my perspective, leading to the previously mentioned inter-annotator disagreement (4.2.2) between me and the project member who first labelled the letter. To ensure a better quality of the labels, a precise definition of where to draw the line between the two keywords "Essen/Trinken" and "Wirtschaft" would be necessary, and how to correctly label letters in which Hilde and Roland discuss food ration cards.

For the second example letter, OBF-420319-002-01, the problem is most likely in the decisions made during the summarization process of the old keywords to concept terms rather. The keyword "Einkauf" ("purchase") can describe shopping for food as much as economic worries in this context, caused by the food ration cards. This demonstrates once more how crude the summarizing process of the keywords was, as it did not further consider what the annotators truly intended. Consequently, important topics can get lost.

To sum up, H3 can be rejected. The trajectory does not provide any information on whether economic worries increased as the war progressed, since the relative keyword frequencies do not show an increase. What can be learned from studying the keyword "Wirtschaft" is that a higher F1-score (for instance, compared to "Luftkrieg") does not necessarily mean a better interpretability of the keyword frequencies. Moreover, a peak in the graph does not necessarily indicate a significant event. The peak of the keyword "Wirtschaft" in 1942 is rather caused by a combination of false positives, individual decisions made by the annotator, the problematic labelling process, and the very broad concept term "Wirtschaft", which unites many different topics, such as hunger and difficulties in the purchasing of food, ration marks, and major economic restructurings, such as in the arms industry.

H4: The keyword "Kulturkontakt" ("cultural contact") temporally

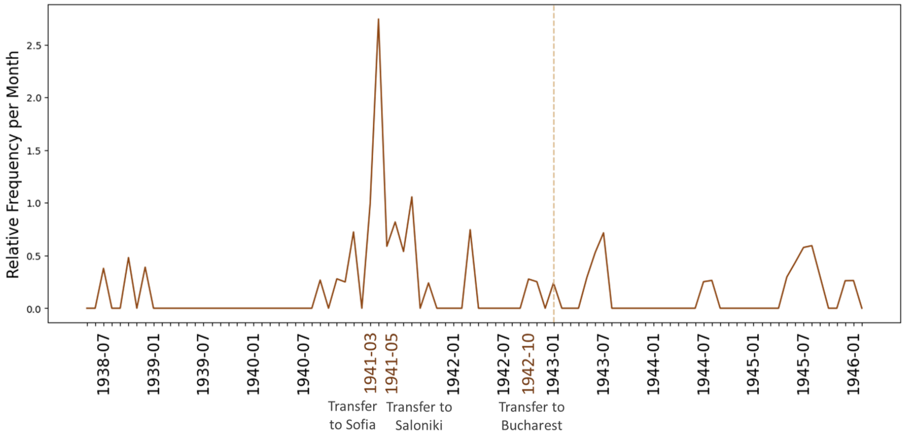


Figure 5.17: Relative Frequency of "Kulturkontakt"

increases each time Roland is newly deployed to a different location during his war service. Hypothesis 4 posits that the keyword "Kulturkontakt" is more frequently used in letters at the outset of new deployments. It is based on the statement of Humburg (2011, 79), who claims that topics in field post letters are often selected based on their novelty. For instance, only at the beginning of a new deployment, a soldier talks about the local population of an occupied country in a derogatory way. Later, he becomes accustomed and used to the new environment and does not feel a need to write about it that much anymore. Hypothesis 4 aims to investigate whether a similar pattern can be observed in Roland's letters.

"Kulturkontakt" is one of the classes that achieves an F1-score of 0 on the test data. Considering that Roland was stationed in Bulgaria, Greece, and Romania before 1943, which is when the data labelled by the classifier begins (dashed line on the graph), it remains valuable to examine the progression curve for the relative frequencies of "Kulturkontakt" before it was predicted by the text classifier (Figure 5.17). Relating the peaks in the graph with Roland's deployments, there is a clear correlation between increased discussion of the topic of cultural contact and Roland's deployment in Bulgaria. The peak in the graph in 1941 indicates that the hypothesis might be correct. Reading the letters from March 1941 clearly shows that here the keyword "Kulturkontakt" was appropriately used that often. Roland explored Sofia and told his wife about his new impressions of the city and the people:

[...] dann sind wir losgebummelt, ohne Plan und Ziel, dicht vorbei am Zigeunerviertel! Daß es da eines gibt ganz in unsrer Nähe, erfuhren wir heute aus einem Befehl, der das Betreten dieses Viertels verbietet.

[...] then we strolled off, without a plan or a destination, right past the gypsy district!

That there is one very close to us, we learned today from an order that forbids the entry into this quarter.
(Fahnenbruck et al. 2023, OBF-410416-001-01)

Hilde, on the other hand, reacted with curiosity:

Sieh Dich nur fleißig um! Du mußt mir erzählen – erzählen, wenn Du heimkommst! Von allem, was Du sahst. Auch von den schönen Dunkeläugigen des Orients! Aber die Luderchen scheinen zum großen Teil gar keine Seltenheit zu sein! Du sagst, sie seien europäisch gekleidet und – bemalt!!

Just look around diligently! You must tell me—tell me when you get home! About everything you saw. Also about the beautiful dark-eyed women of the Orient! But the hussies don't seem to be rare at all for the most part! You say, they are dressed European and—painted!
(Fahnenbruck et al. 2023, OBF-410404-002-01)

The graph shows no peak in 1942. Roland arrived in Bucharest on October 18, 1942 (Fahnenbruck et al. 2023, OBF-421018-001-01). Out of the total of 118 letters from November and December 1942, in which Roland started to explore Bucharest, just as he explored Sofia one and a half years before, only one is labelled with "Kulturkontakt". Upon reading the letters of these months, it is evident that the „Kulturkontakt" label is indeed missing for some, as in the case of the following letter:

Von den Menschen hier kann ich noch nicht viel sagen. Die mir nun so täglich auf meinem Gang begegnen, machen, gemessen an den in S. [Thessaloniki] empfangenen Eindrücken, den Eindruck geringerer Wohlhabenheit, Sauberkeit. Dunkel ist vorherrschend – und ganz allgemein ist wenig Helligkeit und sichtbare Lebensfreudigkeit mit diesen Menschen; kein einheitliches Bild, sondern der Eindruck hochgradiger Vermischung.

I cannot yet say much about the people here. Those I now meet daily on my walk, make the impression of lesser prosperity, cleanliness measured against the impressions received in S. [Thessaloniki]. Darkness is predominant—and in general there is little brightness and visible joy of life with these people; no uniform picture, but the impression of high-grade mingling.
(Fahnenbruck et al. 2023, OBF-421111-001-02)

Roland's focus when strolling through Bucharest was overall not that much on the local population, but he was mainly interested in architecture and buildings, as he realized at some point:

Es wird viel gebaut. Auffällig die mächtigen, im Bau befindlichen Regierungsbauten, Ministerien, ich besinne mich auf nicht weniger als drei. Um einen Flügel des Schlosses selber läuft ein mächtiges Baugerüst. Ja so vielbeschäftigt war das Auge, daß es auf die Menschen kaum geachtet hat.

There is a lot of construction going on. Conspicuous the mighty government buildings under construction, ministries, I recall no less than three. Around one wing of the castle itself runs a mighty scaffolding. Yes, the eye was so busy that it hardly paid attention to the people.
(Fahnenbruck et al. 2023, OBF-421108-001-01)

Technically, there is a keyword “Baukunst” (“architecture/construction art”). If it were properly assigned, one should be able to observe an increase in November 1942, since Roland reports much about the architecture in Bucharest. However, “Baukunst” as a class label reaches an F1-score of 0 on the test data and occurs in the total labelled data only 10 times. Therefore, it can unfortunately not be further studied here.

It has to be concluded for H4 that the frequency of the keyword “Kulturkontakt” can, but not necessarily has to correlate with new deployments. Other factors also play a role, such as Roland’s personal interest in the city of his deployment and its inhabitants.

H5: It can be observed from the letters dealing with the topic of hygiene that ideas of hygiene not only refer to physical cleanliness but also include racist ideas. Hypothesis 5 is based on a study by Kipp (2014). The author worked with 7,000 field post letters to study soldiers’ individual perceptions of the occupied countries where they were deployed (Kipp 2011, 457). The basic civilizational motif of “Reinlichkeit” (“cleanliness”) reflects the extent to which Nazi propaganda affected individuals’ perceptions (Kipp 2011, 464). Hypothesis 5 is therefore a search for traces of the “Reinlichkeit” motif in the Oberfrohna correspondence.

Investigating the hypothesis requires a method more along the lines of Close Reading, as mere trends or peaks in the curve might not provide substantial insights into the actual significance of the term. The curve can, at the very least, function as an indicator pointing towards specific months that warrant closer examination. These months seem to exhibit a heightened relevance of the topic for both Hilde and Roland. Figure 5.18 shows the curve for the keyword “Hygiene”.

Moreover, the corpus can be systematically searched by the buzzwords “Reinlichkeit”, “Sauberkeit” (“cleanliness”), “Reinemachen” (“cleaning up”), “Großreinemachen” (“major cleaning”) “schmutzig” (“dirty”) and “Säuberung” (“cleaning”), which in their basic meaning describe physical hygiene, but were often instrumentalized by the Nazis (Kipp 2011, 467). Searching explicitly for these terms goes beyond the original method of just looking at keyword trends, however, it opens the possibility for finding further letters which might either not have been correctly labelled or have been correctly labelled but were not written within a month with a peak of the “hygiene” topic. Letters on the topic of “hygiene” that additionally contain the terms of interest may also have a higher probability of being at the intersection of physical hygiene and racial hygiene. They might illustrate if the concept of physical hygiene and the violence in the annihilation war were interconnected for Hilde and Roland.

The new concept term “Hygiene” incorporated only old keywords that indeed

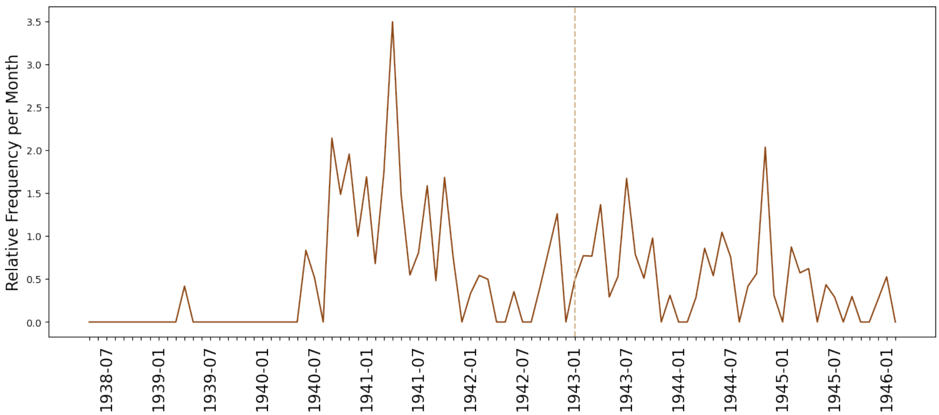


Figure 5.18: Relative Frequency of "Hygiene"

appear to relate to physical hygiene, cleaning, and aesthetics:

- "Bad" ("bath")
- "baden" ("bathing")
- "Dusche" ("shower")
- "duschen" ("showering")
- "Friseur" ("hairdresser")
- "Frisur" ("haircut")
- "Hygiene" ("hygiene")
- "Hygieneartikel" / "-produkte" ("hygiene products")
- "reinigen" / "Reinigung" ("cleaning")
- "Wäsche" ("laundry")
- "waschen" ("washing")

The topic of hygiene served several purposes in many field post letters. It allowed soldiers to communicate their traumatic experiences, brutality, and cruelty without explicitly having to mention them, which would have been inappropriate for men (Kipp 2014, 70). Moreover, the "hygienic" Wehrmacht soldiers distinguished themselves from the "dirty" people in Eastern Europe and identified with the task of "big cleaning" (Kipp 2011, 456). An examination of the letters on the topic "Hygiene" should now show whether such a phenomenon is also evident in Hilde and Roland's correspondence.

As starting points for the examination serve the months of April 1941 (14 out of 55 letters labelled with "Hygiene") and November 1944 (8 out of 51 letters). Closely examining these letters suggests that the letters of April 1941 are labelled correctly and only refer to actual physical hygiene. Out of the 8 letters from November 1944, 4 turn out as false positives, which might in the case of (Fahnenbruck et al. 2023, OBF-441105-001-01, unpublished) for instance be caused by the expression “Es waschen sich zu viel die Hände drin.” (“Too many wash their hands in it.”) Instead of literally washing one’s hands, Roland refers to the budget for food for soldiers in Romania, which is low because others get rich from it. The four true positives from November 1944 refer to physical hygiene.

Searching the corpus for the buzzwords “Reinlichkeit”, “Sauberkeit”, “Großreinemachen”, “schmutzig”, and “Säuberung” primarily shows how for Hilde and Roland cleanliness is first and foremost emblematic of home. This dichotomy of a home that the wife is supposed to keep clean and the dirty front that the husband must heroically endure is a phenomenon that Kipp (2014, 276–77) could observe in the letters they analysed. In the Oberfrohna correspondence, Hilde, the housewife, acts as the manufacturer of the cleanliness at home, as opposed to the dirty front. The need for cleanliness in the foreign country is representative of Roland’s longing for home. Nowhere else it can be as clean as at home with Hilde. They dream of a future in a clean home after Roland’s return. In the following example, Hilde described cleaning her house and afterwards her own physical hygiene in order to be clean for Roland despite the distance:

Nun ist alles blitzblank, kann der liebe Hubo kommen!!!!! Und mir gefällt es nun erst wieder mal richtig bei uns. Ich hab’s [z]u gerne, wenn um mich her peinliche Sauberkeit herrscht, dann erst fühle ich mich zuhaus. Aber nun war es auch Zeit, daß ich mich vom Schmutz befreite, mich persönlich!! Kann doch nicht beschmutzt und im ‚Scheuerstaat‘ mit meinem Herzlieb plaudern! So nahm ich mir einen großen Asch [sic] voll Wasser und seifte mich fein ab – von Kopf bis zu Fuß – ganz im Evakostüm, Du!!

Now everything is spotless, dear Hubo [nickname for Roland] can come!!!!! And eventually I really like our place again. I like it when there’s meticulous cleanliness around me, then I feel at home. But now it was also time that I freed myself from the dirt, me personally! I can’t chat with my Sweetheart when I’m dirty and in the ‘scrub state’! So I took a big bowl of water and soaped myself off finely—from head to foot—completely in the birthday suit, you!!!
(Fahnenbruck et al. 2023, OBF-410529-002-01)

Roland romanticized how their togetherness creates livability and cleanliness:

Herzelein, zum Schloß ist uns schon das kleinste bescheidenste Stübchen geworden, für kurze Zeit freilich. Wir wollen uns auch ein Nestlein bauen nach unserem Sinn. Sauber und ordentlich soll es sein, wohnlich, heimlich, einladend, gemütlich.

Sweetheart, the smallest, most modest room has already become our castle, admittedly for a short time. We also want to build a little nest in our sense. It should be clean and tidy, homely, cozy, inviting, comfortable.
(Fahnenbruck et al. 2023, OBF-430605-001-01, unpublished)

Another point Kipp (2011, 467) raises in this context is positive motivation for soldiers through cleanliness in the barracks. The living conditions in the foreign country are to be adapted to the ones at home in Germany. In some letters, it appears that Roland employed this strategy. Through cleanliness in the barracks, he created a link to home, where Hilde ensured cleanliness and from which he could draw strength. To a letter in which Roland told her about his clean room, Hilde responded with a positive affirmation:

Wie freue ich mich, wenn Du mir erzählst, daß es in Euerm Stübel immer wohnlicher wird, daß eins nach dem andern sich einstellt, bei der Sauberkeit angefangen. So habe ich mir's doch schon lange gewünscht, daß Du es mal triffst beim Militär. Wenn Du aus dem Dienst kommst, möchtest Du einkehren können in einer Bleibe, die Dir dann ein wenig von dem Zauber heimatlicher Traute verkörpert.

How happy I am when you tell me that it is getting more and more livable in your room, that one thing after another sets in, starting with the cleanliness. That's what I've been wishing for a long time, that you're lucky in the military. When you get out of the service, you'd like to be able to stay in a place that embodies a little of the magic of home.
(Fahnenbruck et al. 2023, OBF-421110-002-01)

Moreover, for Hilde and Roland physical hygiene was often intertwined with spiritual and moral cleanliness, which is another typical phenomenon in field post letters: "Alles Saubere und Reine, dessen der Soldat gewahr werden konnte, stand auch – im übertragenen Sinn – für die Heilkraft der Liebe gegenüber den Wunden des Krieges." ("Everything clean and pure, of which the soldier could become aware, also represented—figuratively—the healing power of love opposed to the wounds of war.") (Kipp 2014, 70). In the context of their marriage, Hilde and Roland applied their idea of cleanliness to their character traits and their relationship:

Der Gedanke an Verrat und Treulosigkeit – er liegt ja so ferne. [...] Von Schmutzigen, Häßlichen hebt sich das Reine, Schöne noch einmal so gut ab.

The thought of betrayal and disloyalty—it is so far away. [...] From the dirty and ugly the pure and beautiful stand out even more.
(Fahnenbruck et al. 2023, OBF-420103-002-01)

Furthermore, Kipp (2011, 466) found a lot of evidence for hateful characterizations of the "Dirty East" and "culturally inferior" peoples in field post letters. They claim that for Germans, ideas of cleanliness could serve as a means of distinction and to defame, desexualize, and dehumanize the "dirty" population of the occupied territories. A targeted search for the terms "Reinemachen" ("cleaning") and "Großreinemachen" ("major cleaning") reveals no evidence that ethnic cleansing was a topic of discussion for Hilde and Roland. The terms refer exclusively to cleaning in the house. However, there is a limited number of letters with attempts to demarcate "dirty" Eastern Europe to "clean" Germany. In the following example, Roland told Hilde about some of his impressions from Bucharest. A lack of cleanliness arose not only from an actual lack of hygiene but also the women's red-painted fingernails disgusted Roland. He associated them with dirt:

Aber die Frauensleute, die [im Friseurladen] in der Damenabteilung hantieren, sind nicht eben die saubersten. Deren Händen – mit rotlackierten Nägeln – möchte ich mich nicht anbefehlen – brrr – dabei erscheint ganz ansehnliche Kundschaft. Die Sauberkeit steht eben nicht allzuhoch [sic] im Kurse – Rumänien!

But the women who work in the ladies' section [of the barbershop] are not exactly the cleanest. I wouldn't want to place myself in their hands—with red-painted nails—brrr—even though quite respectable customers appear. The cleanliness is just not too high in the courses—Romania!
(Fahnenbruck et al. 2023, OBF-431002-001-01, unpublished)

Roland and Hilde also expressed their opinions on the population of the hostile Soviet Union in the context of hygiene. Interestingly, on the one hand, Hilde described here the common clichés and one can observe exactly the strategy of distinction from the dirty mentioned by Kipp (2011, 466):

Wenn ich da die alten schmutzigen Russenfrauen täglich vorbeigehen sehe, gruselts mich richtig. Aber die haben gar kein solch Verlangen nach Sauberkeit.

When I see the dirty old Russian women walking by daily, it really disgusts me. But they have no such desire for cleanliness.
(Fahnenbruck et al. 2023, OBF-430108-002-01, unpublished)

Roland, on the other hand, did reflect these prejudices critically, disagreed with them, and even criticized society for its ideas of "dirty" Russians. In his description of the Russians, moral and physical hygiene go hand in hand:

Wir sind doch Deutsche! Ja, ja! Wenn nur in allen Deutschen noch soviel Heimatgefühl, soviel Volksgut lebendig wäre wie in diesen einfachen Menschen aus der Ukraine, aus Sovjetrußland! Und soviel Anständigkeit. Man sagt nämlich den Russen und Russinnen Sauberkeit und moralische Gesundheit nach! Ja, Herzelein! Wann wird der Deutsche diesen blöden Dünkel einmal fahren lassen? Wann wird er lernen, die Eigenart des anderen zu respektieren? Und sich selbst zu erkennen?

We are Germans after all! Yes, yes! If only in all Germans there was still alive so much sense of home, so much national heritage as in these simple people from Ukraine, from Soviet Russia! And so much impudence. Russians are said to be clean and morally healthy! Yes, Sweetheart! When will the German let go of this stupid conceit? When will he learn to respect the peculiarities of others? And to recognize himself?
(Fahnenbruck et al. 2023, OBF-430625-001-01, unpublished)

Also, the idea of the "dirty" Russians and the accompanying self-righteousness of the Germans was for Roland obviously a product of Nazi propaganda:

Ach soviel Irrtum und Verwirrung auch gerade durch das Geschehen der letzten Tage. Die Propaganda hat alle Sinne benebelt, sie hat soviel Selbstgerechtigkeit aufgerichtet *und soviel Kritik und Selbsterkenntnis niedergerissen. Unsre Sache war die gerechte – wie könnte der Sieg an den Feind fallen – an den schmutzigen Russen – es gibt keine Gerechtigkeit – so hört man reden unter Kameraden. Was wissen wir Menschen um die Gerechtigkeit und Weisheit von Gottes Führung?!!!*

Oh so much error and confusion especially by the events of the last days. Propaganda has dulled all senses, it has raised so much self-righteousness and torn down so much criticism and self-reflection. Our cause was the righteous one—how could the victory fall to the enemy—to the dirty Russians—there is no justice—so one hears talk among comrades. What do we humans know about the justice and wisdom of God's guidance?!!!
(Fahnenbruck et al. 2023, OBF-450513-001-01, unpublished)

In the same letter, only a few lines later, there is a clear indication that Roland did not agree with the actions of the Nazis and the idea of Germans as "master race", at least not anymore by the end of the war. He could only guess about what was going on in the concentration camps, but by the end of the war, he learned how many people actually were there in captivity and murdered. He expressed shock at the happenings and the Nazis' overconfidence and admitted to feeling collective guilt:

Mit Vorbehalt ist gewiß alles aufzunehmen, was jetzt von der anderen Seite über den Funk zu uns kommt, aber einige Angaben und Zahlen mögen doch erhärten und verstärken, was wir schon vorher als ein dunkles Kapitel unsrer Schuld, unsrer maßlosen Vermessenheit und Überheblichkeit kannten, ich meine die Angaben über die Opfer der Konzentrationslager. Das ist tiefste Barbarei! Das ist Unmenschlichkeit. Das ist grausame Tyrannei, das ist gottlose Anmaßung eines Richteramtes, das ist tiefste Finsternis – in diesem Zeichen kann das Leben nicht fortgehen!

Everything that reaches us now from the other side via the radio has to be taken with reservations, but some data and figures may confirm and strengthen what we already knew before as a dark chapter of our guilt, our excessive presumption and arrogance, I mean the data about the victims of the concentration camps. This is the deepest barbarism! This is inhumanity. This is cruel tyranny, this is godless arrogance of a judicial office, this is deepest darkness—in this sign, life cannot go on!
(Fahnenbruck et al. 2023, OBF-450513-001-01, unpublished)

Finally, it should be summed up what has been found when studying the topic "hygiene" in the Oberfrohna correspondence, using keyword frequencies as the established method of this thesis, buzzword search for interesting terms in this context, as well as Close Reading of relevant letters. Overall, the investigation has shown that the topic of hygiene in the correspondence carries less racial connotation than expected based on Kipp's (2011; 2014) study. Instead, for the newlyweds hygiene and cleanliness were often symbolical for missing home and being together. The clean home sometimes did stand in contrast to the dirty foreign country, but the emphasis was less on the hostile country (especially Romania) itself being dirty and more on cleanliness as an emblem of security, safety, and their shared married life. Throughout the correspondence, Roland and Hilde said little about the subject of concentration camps and Nazi “racial hygiene”. Only towards the end of the war, Roland made clear that he strongly opposed these ideas.

Chapter 6

Conclusion

The eight-year letter correspondence between Hilde Laube and Roland Nordhoff is a valuable source of firsthand accounts from “common people” in the pre-war and World War II years. It is important to digitize the letters and thus preserve them for future generations of researchers and interested readers. In this thesis, I used these now digitally available letters to train a text classifier and perform analyses using the letters’ thematic keywords.

The thesis first introduced the letters and their writers, Hilde and Roland. Hilde stayed in the village Oberfrohna in rural Saxony, while her husband served as a military scribe in various Eastern European countries and in the later years of the war in Kiel, Germany. Because of his position as a military scribe, Roland could not only often avoid direct front-line combat but also had plenty of time to correspond with his wife back home.

Thus, between May 1938 and February 1946, the two spouses wrote at least 2,626 letters to each other. The number of letters written by each is almost balanced: Hilde wrote 1,128 letters, Roland wrote 1,498 letters. Especially in terms of tokens, the two spouses were nearly equal. Out of the 3,098,969 words the entire correspondence contains, Hilde wrote 1,539,489 and Roland wrote 1,559,480. Some letters of the correspondence got lost either in the mail or after being delivered for unknown reasons. Therefore, we will never know conclusively how many letters the two spouses really wrote.

The letters that did not get lost have been preserved by Hilde and Roland’s family. An international and interdisciplinary team of citizen scientists, humanities scholars, students, and interns now successively digitizes them and publishes them in the context of the public history project *Alltag im Krieg*. As a year-long project member of AiK and its predecessor project *Trug&Schein*, I have close insights into the crowdsourced digitization and publication process of the letters. It was important to me to provide a detailed account of the dedication and efforts of the individuals involved in these processes (especially the project leaders Laura Fahnenbruck and Andrew Bergerson) who contribute almost exclusively on a voluntary basis. This collective effort is what allows for the long-term preservation and accessibility for the public of a corpus of this magnitude.

The corpus I worked with was partially available from the AiK project website. As the digitization and publication of the letters are still ongoing, only about half of them have been made available there. The remaining part of the corpus consists of currently unpublished letters. Most of these are transcriptions generated by the transcription software Transkribus and have not yet undergone proofreading and may still contain transcription errors. The part of the correspondence that has been proofread and published already is also labelled with 81 different thematic keywords. I used this part of the corpus to train various text classifiers and select the best model.

Identifying the most appropriate classifier for the Oberfrohna correspondence was the first part of the research objective. I selected four different models (a rule-based model based on the keywords themselves, Logistic Regression, SVM, and LSTM) to assign multiple keywords to each letter in a test set as accurately as possible. The overall performance of each model, as well as the performance on the 20 most frequent classes, was reported after fine-tuning the parameters. The most efficient model is a combination of Logistic Regression, SVM with sigmoid kernel, and the rule-based classifier. A majority voting of these classifiers results in an overall micro F1-score of 0,34 (micro-precision: 0.29; micro-recall: 0.41) on the test data. The individual F1-scores for each class range from 0 (27 out of 81 classes) to 0.67 (keyword "Gefühle", which is also the most frequent in the corpus).

The second part of the two-part research objective was to inspect the individual keywords closely over time. The research interest focused on both trends and continuities in the relative frequencies of specific keywords. The relative frequency for a keyword per month is the normalized proportion of letters labelled with it. Normalization was achieved by dividing the number of letters labelled with a specific keyword by the logarithm of all letters for that month. Using normalized frequencies aims to address the comparability problem of timelines. The total number of letters available for each month in the correspondence varies strongly within the correspondence, making it challenging to compare the significance of a specific topic between individual months. Normalization can enhance the smoothness of line charts depicting keyword trends over time and improve the comparability of individual months.

The main assumption for investigating keywords was that their frequencies reflect the relevance of a topic for Hilde and Roland. If a particular keyword pertains to a social or political topic, its frequency plot should, therefore, depict real-world events. To confirm any correlation between real-world events and anomalies in the frequency line charts, closely reading the letters from the corresponding months is necessary. To analyse the keywords, I pursued an exploratory approach as well as a narrower focus guided by initially posed hypotheses.

The goal of the exploratory approach was to identify keywords that are particularly outstanding due to unexpected continuities, trends, or peaks in the line charts they form. I used both the original keywords of the letters and the keywords assigned by the classifier for the analysis. Counting keyword frequencies allowed to compare topic focuses between Hilde and Roland. Using the coefficient of variation to identify "standard themes" in the correspondence showed that Hilde and Roland wrote about the topics "Arbeit", "Mobilität", "Gefühle", "Familie", and "Schreiben" relatively consistently often over the years. The topics "Essen/Trinken" and "Kameraden" appear to gain relevance in the correspondence only after Roland's enlisting in military training in August 1940. The keyword "Führer" turned out to distinctively reflect the 10^{th} anniversary of Hitler's rise to power, and, to a lesser extent, Hitler's death. However, this approach leaves a lot of freedom and allows to analyse only those keywords which appear to be interesting based on their frequency over time.

The approach with a narrower focus examined hypotheses which had been formulated *before* assigning the keywords with the classifier, and therefore before it was known how frequent the keywords would be over the years of the correspondence and how the line charts would look like. This approach allows us to better investigate the extent to which analysing keyword frequencies over time is a suitable method to provide information about the corpus on a wide variety of topics:

- Hypothesis 1 on correlations of the keyword “Luftkrieg” (“air war”) with actual air raids could be partially confirmed since the line chart reflects air raids of the British in 1940 as well as in 1945. However, this does not necessarily imply that these two corresponding air raids were the most relevant ones to Hilde and Roland. Others that are equally relevant in the correspondence might not be as strongly reflected by the keyword. (However, it can be assumed that for people from Saxony, the bombing in 1945 was one of the most traumatic events and therefore justifiably appears most prominently in the graph.)
- Hypothesis 2, which assumes that topics related to culture might become less frequent in the letters towards the end of the war, had to be rejected. The graph does not show any indications of a phenomenon like this.
- Hypothesis 3, about an increase of the keyword “Wirtschaft” towards the end of the war, had to be rejected, too. However, a close analysis showed that the keyword is somewhat problematic and would need to be more precisely distinguished from “Essen/Trinken” in order to better analyse it.
- Hypothesis 4 states that the keyword “Kulturkontakt” is more frequent at the beginning of new deployments, when Roland told his wife about his first encounters with the local population. The graph does show an outstanding increase in 1941 (deployment in Greece and Bulgaria), but not in 1942 (deployment in Romania). The conclusion is, therefore, that the hypothesis is not universally valid, and an intensive reporting of contact with the population of the occupied country also depends on other factors.
- Hypothesis 5 aimed to explore the relationship between physical hygiene and the concept of racial hygiene in Nazi Germany. A close examination of the topic of “Hygiene” showed that Hilde and Roland did not align with the Nazi ideology of racial hygiene. For them, hygiene rather served to separate their own “clean” marriage from the “dirty” outside world. Thus, the hypothesis needed to be discarded, yet it led to an alternative discovery.

The thesis excessively discussed the limitations of the applied methods. Training the classifier for assigning the keywords turned out as a demanding task. Especially for a neural network, a dataset with only 728 documents is not sufficiently large to achieve a good performance. The performance of all models for very infrequent classes can hardly be evaluated on a test set with only 264 documents. Also, the class labels used for training and evaluation are noisy, making it difficult for the model to clearly distinguish between the classes. Therefore, the automatically assigned keywords cannot be used when publishing the letters, as I originally intended.

Interpreting the original and the obtained keywords requires a high amount of caution. On the one hand, in many cases, these keywords are not correct or not complete, as a closer examination of the letters' content revealed, and a high level of randomness must be considered when analysing them. On the other hand, months with missing letters or periods of, for instance, furloughs result in poorer comparability between the individual months. What Hilde and Roland discussed during these periods remains forever unknown. It is therefore problematic to aim for detailed information about specific topics in Hilde and Roland's correspondence only from the keywords. However, it is possible to gain a general overview of the corpus. The keywords do give readers information about which topics were more important than others and how the topic focuses between the two spouses differed. However, only closely reading specific letters from certain periods can really enhance our knowledge about the role of individual topics in the correspondence.

Overall, examining the hypotheses and reading selected letters suggests that a temporarily higher relative keyword frequency generally corresponds to a higher importance of a topic for Hilde and Roland. However, this does not rule out the possibility that there might be other periods when the topic was equally or even more important, but the letters were not labelled correctly either by the classifier or by the annotators. In addition, it is possible that annotators concentrated heavily on one topic for a specific period and labelled many letters with it, even though compared to other periods of the correspondence, the topic does not occur significantly more in the letters.

In an outlook for the future, we can expect an improvement in overall data quality as the AiK project continues to evolve. Repeating the thesis' analyses in a later project stage could yield improved and clearer results. The meanings of keywords have been clearly defined since the transition from T&S to AiK, which will provide more consistent class labels for all letters labelled from now on. These improvements would simplify classifier training. However, it is evident that training a classifier with cleaner data is no longer necessary once project members have manually annotated the complete corpus with correct labels. One potential approach, if project resources permit, would involve correcting the existing AiK labels which were initially generated through the problematic summarization process of the old keywords in T&S, and then rerunning the training with the corrected labels. I decided against re-labelling the 1,304 previously labelled letters for the analyses in the master's thesis because of the enormous amount of work involved. Now, after running all experiments and analyses, however, I must admit that it would probably have been worth the time. Most likely, it would have resulted in better and less ambiguous results than treating the existing class labels despite their low quality as ground-truth when training the models.

For certain keywords (especially "Führer" and "Luftkrieg"), the created line charts were indeed able to provide valuable information about the correspondence. This demonstrates that the method in general appears to work. We can assume that the line charts for some other keywords would also be more informative at a later stage of the AiK project after the corpus has been correctly labelled by human annotators. I will, therefore, definitely recreate all line charts for further analysis once the AiK project is completed and all letters have been proofread, labelled, and published by the project members.

For future work with the AiK corpus, whether applying methods of digital history or "only" traditional history, the topic of hygiene appears to have great potential. Out of all the keywords studied, "Hygiene" yielded the most surprising results and could offer the deepest insights into the correspondence, albeit not entirely aligned with the initial hypothesis. It must be taken into account that I searched the corpus for this topic slightly more extensively. In addition to the letters labelled with the keyword "Hygiene", I explored letters containing specific buzzwords related to the topic. Nevertheless, the study of the topic was relatively brief. Still, it could reveal quickly how hygiene transcends the concept of physical cleanliness, extending to a symbolic level. This initial analysis can serve as a foundation for further exploration.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al., ‘TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems’, 2015 <https://www.tensorflow.org/>

Agarwal, Sumeet, Shantanu Gudbole, Shourya Roy, and Diwakar Punjani, ‘How Much Noise Is Too Much: A Study in Automatic Text Classification’, in Seventh IEEE International Conference on Data Mining (presented at the ICDM 2007, Omaha, 2007), pp. 3–12 <https://doi.org/10.1109/ICDM.2007.21>

Aggarwal, Charu C., ‘An Introduction to Data Classification’, in Data Classification. Algorithms and Applications, ed. by Charu C. Aggarwal (Boca Raton: CRC Press, 2015), pp. 1–36

———, ‘Instance-Based Learning: A Survey’, in Data Classification. Algorithms and Applications, ed. by Charu C. Aggarwal (Boca Raton: CRC Press, 2015), pp. 157–85

Aggarwal, Charu C., and ChengXiang Zhai, ‘Text Classification’, in Data Classification. Algorithms and Applications, ed. by Charu C. Aggarwal (Boca Raton: CRC Press, 2015), pp. 287–336

Apostolova, Emilia, and R. Andrew Kreek, ‘Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data’, in Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-Generated Text, ed. by Association for Computational Linguistics (presented at the 2018 EMNLP Workshop W-NUT, Brussels, 2018), pp. 104–9

Arhin, Kofi, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh, ‘Ground-Truth, Whose Truth? - Examining the Challenges with Annotating Toxic Text Datasets’ (arxiv, 2021) <https://arxiv.org/abs/2112.03529>

Au Yeung, Ching-man, and Adam Jatowt, ‘Studying How the Past Is Remembered: Towards Computational History through Large Scale Text Mining’, in CIKM ’11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (presented at the CIKM ’11, Glasgow, 2011), pp. 1231–40 <https://doi.org/10.1145/2063576.2063755>

Baker, K. Scott, Andrew S. Bergerson, Laura Fahnenbruck, Deborah Parker, and Benjamin Roers, ‘The Ongoing Rewards of Collaboration, Intermediality, and Multivocality in the Humanities: Reflections on the Multimedia Project *Trug&Schein*’, in Feminist German Studies, ed. by Hester Baer and Alexandra M. Hill, Feminist German Studies, 36 (Lincoln: University of Nebraska Press, 2020), I, 54–70

Beltagy, Iz, Matthew E. Peters, and Arman Cohan, ‘Longformer: The Long-Document Transformer’ (arxiv, 2020) <https://arxiv.org/abs/2004.05150>

Bergerson, Andrew S., ‘Das Sich-Einschreiben in die NS-Zukunft. Liebesbriefe als Quelle für eine Alltagsgeschichte der ›Volksgemeinschaft‹’, in Der Ort der ›Volksgemeinschaft‹ in der deutschen Gesellschaftsgeschichte, ed. by Detlef Schmiechen-Ackermann, Marlis Buchholz, Bianca Roitsch, and Christiane Schröder, Nationalsozialistische ›Volksgemeinschaft‹. Studien zu Konstruktion, gesellschaftlicher Wirkungsmacht und Erinnerung (Leiden: Ferdinand Schöningh, 2018), Band 7, 223–41

Bergerson, Andrew S., and Laura Fahnenbruck, ‘Archivierung des Projektes Trug&Schein: Ein Briefwechsel. Eine kritische Begegnung mit dem Alltag des Zweiten Weltkriegs—Schreib mit! (2011-2022) Archivierung: 08/2022’, 2022 <https://alltag-im-krieg.de/sites/default/files/imce-media/PDF/OBF%20Vorg%C3%A4ngerprojekt%20T%26S_barrierefrei.pdf> [accessed 29 April 2023]

Bergerson, Andrew S., Laura Fahnenbruck, and Christine Hartig, ‘Working on the Relationship. Exchanging Letters, Goods and Photographs in Wartime’, in Private Life and Privacy in Nazi Germany, ed. by Elizabeth Harvey and others (Cambridge: Cambridge University Press, 2019), pp. 256–79

Bergerson, Andrew S., Thomas Muntschick, and Johannes Schwartz, ‘Trug und Schein: Ein Briefwechsel | Clio-online’, Clio-online <https://www.clio-online.de/webresource/id/webresource-27117?language=en> [accessed 29 April 2023]

Bi, Yaxin, David Bell, Hui Wang, Gongde Guo, and Kieran Greer, ‘Combining Multiple Classifiers Using Dempster’s Rule of Combination for Text Categorization’, in Modeling Decisions for Artificial Intelligence: First International Conference, MDAI 2004, Barcelona, Spain, August 2-4, 2004, Proceedings, ed. by Vicenc Torra and Yasuo Narukawa, Lectore Notes in Artificial Intelligence | Subseries of Lecture Notes in Computer Science, 2004, pp. 127–38 <https://doi.org/10.1080/08839510601170887>

Breidbach, Olaf, and Thomas Bach, eds., Ernst Haeckel Online Briefedition (Jena: Friedrich-Schiller-Universität, 2023) <https://haeckel-briefwechsel-projekt.uni-jena.de/de> [accessed 9 June 2023]

Buchbender, Ortwin, 'Zur Geschichte der Rezeption und Sammlung von Feldpost in Deutschland oder»Auf den Spuren von subjektiven Wirklichkeiten«', in Schreiben im Krieg – Schreiben vom Krieg: Feldpost im Zeitalter der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 17–21

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos, 'Large-Scale Multi-Label Text Classification on EU Legislation' (arxiv, 2019) <https://arxiv.org/abs/1906.02192>

Cohen, Daniel J., and Roy Rosenzweig, eds., Digital History. A Guide to Gathering, Preserving, and Presenting the Past on the Web (Philadelphia: University of Pennsylvania Press, 2005)

Curators of the University of Missouri, 'Andrew Bergerson | Humanities and Social Sciences', University of Missouri - Kansas City, 2023 <https://shss.umkc.edu/profiles/history/andrew-bergerson.html> [accessed 8 May 2023]

Didczuneit, Veit, and Gunnar Goehle, 'Die Briefsammlungen der Museumsstiftung Post und Telekommunikation', Museumsstiftung Post und Telekommunikation Briefsammlung <https://www.briefsammlung.de/> [accessed 29 May 2023]

Durant, Alan, '"The Significance Is in the Selection": Identifying Contemporary Keywords (Final Accepted Version (with Author's Formatting))', 2008 <https://eprints.mdx.ac.uk/8169/>

Dwyer, Larry, 'Thomas-Morse S-4 Scout', The Aviation History Online Museum, 2013 <http://www.aviation-history.com/thomas/scout.html> [accessed 15 July 2023]

Elmalech, Avshalom, and Barbara J. Grosz, '"But You Promised": Methods to Improve Crowd Engagement In Non-Ground Truth Tasks', in Proceedings of the Fifth Conference on Human Computation and Crowdsourcing (HCOMP 2017) (presented at the HCOMP 2017, Quebec City, 2017), pp. 21–30

Ezen-Can, Aysu, 'A Comparison of LSTM and BERT for Small Corpus' (arxiv, 2020) <https://arxiv.org/pdf/2009.05451.pdf>

Fahnenbruck, Laura, and Andrew S. Bergerson, '[OBF] Ausführliche Darstellung der Überlieferung und (ortho-)graphischer Besonderheiten', 2023 <https://alltag-im-krieg.de/sites/default/files/imce-media/PDF/OBF%20Vorg%C3%A4ngerprojekt%20T%26S_barrierefrei.pdf>

Fahnenbruck, Laura, Andrew S. Bergerson, Caroline Bergen, Anne Quiring, and Christian Römmer, 'Oberfrohna', Alltag Im Krieg, 2023 <https://alltag-im-krieg.de/cid/oberfrohna-obf-1> [accessed 13 May 2023]

Feng, Shuo, Huiyu Zhou, and Dong Hongbiao, 'Using Deep Neural Network with Small Dataset to Predict Material Defects', Material and Design, 162, 2019, 200–310

Fridlung, Mats, and Rene Brauer, 'Historicizing Topic Models. A Distant Reading of Topic Modeling Texts within Historical Studies', in Cultural Research in the Context of "Digital Humanities": Proceedings of International Conference 3-5 October 2013 (presented at the Cultural Research in the Context of "Digital Humanities", St. Petersburg, 2013), pp. 152–63 <https://www.researchgate.net/publication/261703677_Historizing_topic_models_A_distant_reading_of_topic_modeling_texts_within_historical_studies>

Fürber, Christian, Data Quality Management with Semantic Technologies (Wiesbaden: Springer Fachmedien Wiesbaden, 2016)

Genkin, Alexander, David D. Lewis, and David Madigan, 'Large-Scale Bayesian Logistic Regression for Text Categorization', Technometrics, 49.3 (2007), 291–304 <https://doi.org/10.1198/004017007000000245>

Gunasekara, Isuru, and Isar Nejadgholi, 'A Review of Standard Text Classification Practices for Multi-Label Toxicity Identification of Online Content', in Proceedings of the Second Workshop on Abusive Language Online (ALW2) (presented at the ALW2, Brussels, 2018), pp. 21–25

Hämmerle, Christa, '"Waiting Longingly ..." Love Letters in World War I. A Plea for a Broader Genre Concept', Max Planck Insitute for Human Development, 2023 <https://www.history-of-emotions.mpg.de/texts/love-letters-in-world-war-one> [accessed 8 June 2023]

Hamp, Birgit, and Helmut Feldweg, 'GermaNet - a Lexical-Semantic Net for German', in Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (presented at the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, 1997)

Henrich, Verena, and Erhard Hinrichs, 'GernEdiT - The GermaNet Editing Tool', in Proceedings of the Seventh Conference on International Language Resources and Evaluation (presented at the LREC 2010, Valetta, Malta, 2010), pp. 2228–35 <http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf>

Heuser, Ryan, and Long Le-Khac, 'A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method', Pamphlets of the Standford Literary Lab, 4 (2012) <https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf>

Hochreiter, Sepp, and Jürgen Schmidhuber, 'Long Short-Term Memory', Neural Computation, 9.8 (1997), 1735–80 <https://doi.org/10.1162/neco.1997.9.8.1735>

Huang, Shih-Wen, and Wai-Tat Fu, 'Don't Hide in the Crowd! Increasing Social Transparency Between Peer Workers Improves Crowdsourcing Outcomes', in CHI '13 (presented at the CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, 2013), pp. 621–30

Humburg, Martin, 'Die Bedeutung der Feldpost für die Soldaten in Stalingrad', in Stalingrad. Mythos und Wirkung einer Schlacht, ed. by Wolfram Wette and Gerd R. Ueberschär, Die Zeit des Nationalsozialismus. Eine Buchreihe, 1992, pp. 68–79

———, '»Jedes Wort ist falsch und wahr - das ist das Wesen des Worts. «Vom Schreiben und Schweigen in der Feldpost', in Schreiben im Krieg – Schreiben vom Krieg: Feldpost im Zeitalter der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 75–86

van Hyning, Victoria, 'Harnessing Crowdsouring for Scholarly and GLAM Purposes', Literature Compass, Wiley Online Library, 16 (2019) <https://doi.org/10.1111/lic3.12507>

Ikonomakis, M., S. Kotsiantis, and V. Tampakas, 'Text Classification Using Machine Learning Techniques', WSEAS Transactions on Computers, 8.4 (2005), 966–74

Indukaev, Andrey, 'Studying Ideational Change in Russian Politics with Topic Models and Word Embeddings', in The Palgrave Handbook of Digital Russia Studies, ed. by D. Gritsenko, M. Wijermars, and M. Kopotev (Palgrave Macmillan, 2020), pp. 334–464

Jander, Thomas, 'Gefährliche Worte. Dissens Und Desertion in Kriegbriefen Deutscher Soldaten', in Schreiben Im Krieg – Schreiben Vom Krieg: Feldpost Im Zeitalter Der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 441–55

Jockers, Matthew Lee, Macroanalysis: Digital Methods and Literary History (Champaign, Illinois: University of Illinois Press, 2013)

Jurafsky, Daniel, and James H. Martin, 'Logistic Regression', Speech and Language Processing, 2023 <https://web.stanford.edu/~jurafsky/slp3/>

Keith, Kathrine A., Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor, 'Identifying Civilians Killed by Police with Distantly Supervised Entity-Event Extraction' (Massachusetts Amherst: arxiv, 2017) <https://arxiv.org/abs/1707.07086>

Kingma, Diederik P., and Jimmy Lei Ba, 'Adam: A Method for Stochastic Optimization' (arxiv, 2015) <https://arxiv.org/pdf/1412.6980.pdf>

Kipp, Michaela, »Großreinemachen im Osten«. Sauberkeitsvorstellungen und Feindbilder im NS-Vernichtungskrieg in Osteuropa 1939-1944/45 (Frankfurt/New York: Campus Verlag, 2014)

———, 'Reinlichkeitsvorstellungen in Feldpostbriefen. Herausforderungen Für Die Kriegsgeschichte', in Schreiben Im Krieg – Schreiben Vom Krieg: Feldpost Im Zeitalter Der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 457–68

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, et al., 'Text Classification Algorithms: A Survey', Information, 10.4 (2019) <https://doi.org/10.3390>

Lee, Changsoo, 'How Are "Immigrant Workers" Represented in Korean News Reporting?—A Text Mining Approach to Critical Discourse Analysis', Digital Scholarship in the Humanities, 34.1 (2019), 82–99 <https://doi.org/10.1093/llc/fqy017>

Leung, Kenneth, 'Micro, Macro & Weighted Averages of F1 Score, Clearly Explained | by Kenneth Leung | Towards Data Science', Medium, 2022 <https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f> [accessed 11 June 2023]

Lewis, D. D., R. Ghani, D. Mladenic, and I. Moulinier, 'Workshop Proceedings', in 3rd Workshop on Operational Text Classification (OTC), in conjunction with SIGKDD (Washington DC, 2003)

Malato, Gianluca, 'Are you still using 0.5 as a threshold?', Your Data Teacher, 2021 <https://www.yourdatateacher.com/2021/06/14/are-you-still-using-0-5-as-a-threshold/> [accessed 30 September 2023]

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze, An Introduction to Information Retrieval (Cambridge: Cambridge University Press, 2009)

Marjanen, Jani, Elaine Zosa, Simon Hengchen, Lidia Pivovarova, and Mikko Tolonen, 'Topic Modelling Discourse Dynamics in Historical Newspapers' (Helsiniki, Goethenburg: arxiv, 2020) <https://arxiv.org/abs/2011.10428>

Mayer, Sandra, Peter Andorfer, Timo Frühwirth, Daniel Stoxreiter, Dimitra Grigoriou, and Sean Quigley, eds., Auden Musulin Papers: A Digital Edition (Vienna: ACDH-CH OEAW, 2023) <https://amp.acdh.oeaw.ac.at/index.html> [accessed 9 June 2023]

Miller, George A., 'WordNet: A Lexical Database for English', Communications of the ACM, 28.11 (1995), 39–41

Moeller, Katrin, 'Die Modellierung des zeitlichen Vergleichs als Kernkompetenz von Digital History? Fachliche Voraussetzungen und digitale Konzepte', in Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft, ed. by Karoline Döring, Stefan Haas, Mareike König, and Jörg Wettlaufer, Studies in Digital History and Hermeneutics, 6 (Oldenburg: De Gruyter, 2022), pp. 83–105

Mohammad, Fahim, ‘Is Preprocessing of Text Really Worth Your Time for Toxic Comment Classification?’ (arxiv, 2018) <https://arxiv.org/abs/1806.02908>

Montani, Ines, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, and others, ‘spaCy: Industrial-strength Natural Language Processing in Python’, 2020 <https://doi.org/10.5281/zenodo.1212303>

Moretti, Franco, ‘Conjectures on World Literature’, in Distant Reading, by Franco Moretti (London; New York: Verso, 2013), pp. 43–62

Muehlberger, Guenter, Louise Seaward, Melissa Terras, et al., ‘Transformating Scholarship in the Archives Though Handwritten Text Recognition’, Journal of Documentation, Vol. 75.5 (2019), 954–76 <https://doi.org/10.1108/JD-07-2018-0114>

Müller, Martin Anton, Gerd-Hermann Susen, and Laura Untner, eds., Arthur Schnitzler. Briefwechsel Mit Autorinnen Und Autoren (Vienna: ACDH-CH OEAW, 2018) <https://schnitzler-briefe.acdh.oeaw.ac.at/> [accessed 9 June 2023]

Nagpal, Anuja, ‘L1 and L2 Regularization Methods’, Medium, 2017 <https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c> [accessed 9 August 2023]

Opryshko, Tetiana, and Serhii Nazarovets, ‘Case Study: Citizen Science in Digital Humanities Context’, in DHW ’21: Digital Humanities Workshop. Proceedings (presented at the DHW ’21: Digital Humanities Workshop, December 23, 2021, Kyiv, Ukraine, New York: ACM, 2021), pp. 198–203 <https://doi.org/10.1145/3526242.3526246>

Owens, Trevor, ‘Making Crowdsouring Compatible with the Missions and Values of Cultural Heritage and Organisations’, in Crowdsouring Our Cultural Heritage, ed. by Mia Ridge (Burlington: Ashgate, 2014), pp. 269–80

Panthöfer, Sven, ‘Rumänien Als Verbündeter Des Deutschen Reiches’, LEMO Lebendiges Museum Online, 2015 <https://www.dhm.de/lemo/kapitel/der-zweite-weltkrieg/kriegsverlauf/rumaenien-als-verbuendeter-des-deutschen-reiches.html> [accessed 9 August 2023]

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and others, ‘Scikit-Learn: Machine Learning in Python’, 2023

Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi, ‘Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections’, in Proceedings of the 17th International Conference on World Wide Web, WWW 2008 (presented at the WWW 2008, Beijing, 2008), pp. 91–100 <https://doi.org/10.1145/1367497.1367510>

Roth, Benjamin and Renato Rocha Souza, ’VU Practical Machine Learning for Natural Language Processing’ (lecture and course offered at University of Vienna, summer term 2021) <https://ufind.univie.ac.at/en/course.html?lv=136040&semester=2021> [accessed 30 September 2023]

Scherstjanoi, Elke, ‘Als Quelle nicht überfordern! Zu Besonderheiten und Grenzen der wissenschaftlichen Nutzung von Feldpostbriefen in der (Zeit-)Geschichte’, in Schreiben im Krieg – Schreiben vom Krieg: Feldpost im Zeitalter der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 117–25

Schwamborn, Vera, ‘MaHiPo - Vom Worte-Schlagen durch den Texte-Dschungel | Flusser Studies’, Flusser Studies, 24.Special Issue: Archiving Flusser (2017) <https://flusserstudies.net/node/675> [accessed 8 June 2023]

Schwender, Clemens, ‘Feldpost Als Medium Sozialer Kommunikation’, in Schreiben Im Krieg – Schreiben Vom Krieg: Feldpost Im Zeitalter Der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen, 2011), pp. 127–38

———, ‘Letters Between Home and the Front. Expressions of Love in World War II “Feldpost” Letters’, in Communication of Love. Mediatized Intimacy from Love Letters to SMS. Interdisciplinary and Historical Studies, ed. by Eva Lia Wyss (Bielefeld: transcript, 2014), pp. 213–36

Scott, Sam, and Stan Matwin, ‘Feature Engineering for Text Classification’, in ICML ’99, 1999 <https://www.researchgate.net/publication/2431220_Feature_Engineering_for_Text_Classification>

Scriba, Arnulf, ‘Bombardierung von Dresden 1945’, LEMO Lebendiges Museum Online, 2022 <https://www.dhm.de/lemo/kapitel/der-zweite-weltkrieg/kriegsverlauf/bombardierung-von-dresden-1945.html> [accessed 15 July 2023]

———, ‘Die “Luftschlacht um England”’, LEMO Lebendiges Museum Online, 2015 <https://www.dhm.de/lemo/kapitel/der-zweite-weltkrieg/kriegsverlauf/luftschlacht-um-england-194041.html> [accessed 15 July 2023]

———, ‘Die NS-Volksgemeinschaft’, LEMO Lebendiges Museum Online, 2014 <https://www.dhm.de/lemo/kapitel/ns-regime/innenpolitik/volksgemeinschaft.html> [accessed 26 May 2023]

Sebastiani, Fabrizio, ‘Machine Learning in Automated Text Categorization’ (arxiv, 2001) <https://arxiv.org/abs/cs/0110053>

Shah, Kanish, Henil Patel, Devanshi Sanghvi, and Manan Shah, ‘A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification’, Augmented Human Research, 5.12 (2020) <https://doi.org/10.1007/s41133-020-00032-0>

Smolarski, René, Hendrikje Carius, and Marcus Plaul, ‘Perspectives and Challenges of Historical Research with Citizen Participation: A Critical Reflection on the Example of “Cinema in the GDR”’, in Digital History. Konzepte, Methoden Und Kritiken Digitaler Geschichtswissenschaft, ed. by Karoline Döring, Stefan Haas, Mareike König, and Jörg Wettlaufer, Studies in Digital History and Hermeneutics, 6 (Oldenburg: De Gruyter, 2022), pp. 303–17

Stader, Ingo, ‘Feldpostbriefe – Eine Art „Social Media“ Im Dritten Reich?’, in Schreiben Im Krieg – Schreiben Vom Krieg: Feldpost Im Zeitalter Der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 139–49

Steele, Helen, ‘»Schreiben Oder Schweigen?«Feldpost and Frauenalltag in Vienna, 1943-1945’, in Schreiben Im Krieg – Schreiben Vom Krieg: Feldpost Im Zeitalter Der Weltkriege, ed. by Veit Didczuneit, Jens Ebert, and Thomas Jander (Essen: Klartext Verlag, 2011), pp. 273–81

Strobel, Jochen, ‘Eine Digitale Edition Der Korrespondenz August Wilhelm Schlegels’, in Weimarer Tagung Briefedition Im Digitalen Zeitalter (presented at the Weimarer Tagung Briefedition im digitalen Zeitalter, Weimar, 2011) <https://edoc.hu-berlin.de/bitstream/handle/18452/6634/2-de2.pdf?sequence=1> [accessed 8 June 2023]

Strobel, Jochen, and Claudia Bamberg, eds., August Wilhelm Schlegel: Digitale Edition der Korrespondenz (Dresden, Trier, Marburg: Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB), ompetenzzentrum – Trier Center for Digital Humanities, Philipps-Universität Marburg, 2014) <https://august-wilhelm-schlegel.de/briefedigital/> [accessed 9 June 2023]

Suissa, Omri, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet, ‘Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science’, Association for Information Science and Technology, 73, 2021, 268–87 <https://doi.org/10.1002/asi.24544>

Süß, Dietmar, Tod aus der Luft. Kriegsgesellschaft und Luftkrieg in Deutschland und England (Munich: Siedler, 2011)

Tang, Jiliang, Salem Alelyani, and Huan Liu, ‘Feature Selection for Classification: A Review’, in Data Classification. Algorithms and Applications, ed. by Charu C. Aggarwal (Boca Raton: CRC Press, 2015), pp. 37–64

Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas, ‘A Review of Multi-Label Classification Methods’, International Journal of Data Warehousing and Mining, 3.3 (2006), 1–13 <https://doi.org/10.4018/jdwm.2007070101>

Vaswani, Anish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and others, ‘Attention Is All You Need’ (arxiv, 2017) <https://arxiv.org/abs/1706.03762>

Vinciarelli, Alessandro, ‘Noisy Text Categorization’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27.12 (2005), 1182–1895

Viola, Lorella, and Jaap Verheul, ‘Mining Ethnicity: Discourse-Driven Topic Modelling of Immigrant Discourses in the USA, 1898–1920’, Digital Scholarship in the Humanities, 35.4 (2020), 921–43 <https://doi.org/10.1093/llc/fqz068>

Wang, Po-Wei, and Chih-Jen Lin, ‘Support Vector Machines’, in Data Classification. Algorithms and Applications, ed. by Charu C. Aggarwal (Boca Raton: CRC Press, 2015), pp. 187–204

Williams, Raymond, ‘Introduction’, in Keywords: A Vocabulary of Culture and Society (Oxford: Oxford University Press, 1985), pp. 12–22

Yang, Yiming, ‘An Evaluation of Statistical Approaches to Text Categorization’, Information Retrieval, 1 (1999), 69–90

Yu, Bei, ‘An Evaluation of Text Classification Methods for Literary Study’, Literary and Linguistic Computing, 23.3 (2008), 327–43 <https://doi.org/10.1093/llc/fqn015>

Appendices

Appendix A

Keywords

A.1 Old and New Keywords

Old Keywords	New Keyword
Empfohlen von FAG, Unvollständig	Editorial Remark
Afrika, Algerien, Amerika, Atlantik, Australien, Balkan, Belgien, Böhmen, Breslau, Bukarest, Bulgaren, Bulgarien, Dänemark, Deutschland, England, Europa, Frankreich, Gibraltar, Griechenland, Holland, Iran, Island, Italien, Japan, Jugoslawien, Jugoslawien, Kanada, Karpaten, Kiel, Köln, Korinth, Kreta, Krim, Kroatien, Leipzig, Litauen, Mazedonien, Nordafrika, Osteuropa, Ostsee, Polen, Rhein, Rotes Meer, Rumänien, Russland, Sächsische Schweiz, Schottland, Schyga, Serbien, Smolensk, Sowjetunion, Spanien, Stalingrad, Suezkanal, Tuerkei, Tunesien, Ukraine, Ungarn, Vereinigten Staaten, Versailles, Warthegau, Wien, Wolhynien	Place Name
Achsenmächte, Alliierte, Dreimächtepakt	Allianzen
Alter, Altersunterschied, Alterung, Generation, Generationen, Generationsunterschied, Jugend, Reife	Alter
Antisemitismus, Juden, Reichspogromnacht	Antisemitismus
Anwerbung, Arbeit, Arbeitsalltag, ArbeitsKalender, Arbeitsmarkt, Arbeitspause, Arbeitsräume, Arbeitsstunden, Arbeitstätigkeiten, Aufgabe, Aufgaben, Bahnhofsdienst, Beförderung, Bereitschaft, Bereitschaftsorder, Beruf, Bewährungsbatallion, Büroarbeit, Chef, deutsche Wertarbeit, Dienst, Dienstalltag, Dienstbereitschaft, Dienstfrei, Dienstpflicht, Dienstplan, Dienstverpflichtung, Ehrenamt, Einberufung, Einberufungen, Entlassung, Fabrik, Fabrikarbeit, Karriere, Kriegsdienst, Kriegsdienste, Lehrberuf, Lehren, Lehrer, Lehrer/Schüler, Lehrerberuf, Schaffen, Schreiber, Wachdienst, Wache, Kinderschar, Schar, Schreibstube	Arbeit
Ausbildung, Belehrung, Benotung, Bildung, Erziehung, Exerzieren, Fortbildung, Kampfspiele, Kurs, Marsch, Offiziersausbildung, Rekrutenzeit, Sanitaetskurs, Schießübung, Schule, Schuljahr, Studium, Zielschießen	Aus-/Bildung
Ariernachweis, Ausweis, Gesundheitsbescheinigung, Urkunden, Zeugnis	Ausweise
Architekt, Architektur, Baukunst, Monument	Baukunst
Besuch, Besucher, Einladung, Treffen	Begegnungen
Amtsmiene, Autorität, Beamtentum, Behörde, Bürokratie, Erlaubnis, Gesundheitskontrolle, Höherer Dienst, Honoratioren, Kontrolle, Marinegericht, Militärgericht, Militärische Formulare, Polizei, Sperrstunde, Strafe, Straftaten, Taten, Verbot, Verwaltung, Verweigerung, Vollmacht, Zensuren	Behörde
Bekannte, Gast, Gäste, Kollegen, Kontakte, Mitarbeiter, Mitbewohner, Nachbar, Nachbarn	Bekannte
Haarmode, Kleider, Kleidung, Kopfbedeckung, Schminken, Schmuck, Schuhe, Stiefel	Bekleidung
Bevölkerung, Bevölkerung, Bewohner	Bevölkerung

Old Keywords	New Keyword
Abendbrot, Alkohol, Brot, Ernährung, Essen, Essgewohnheiten, Früchte, Gänsebraten, Gebäck, Getränke, Hausrezept, Kriegskuchen, Kriegsrezept, Kuchen, Nahrungsmittel, Obst, Proviant, Rauchen, Restaurant, Speisen, Süßigkeiten, Tabak, Tee, Trinken, Wein, Zigarre, Zucker, Betrunkenheit	Essen/Trinken
Bruder, Cousin, Ehe, Ehebruch, Ehestreit, Eltern, Elternschaft, Familie, Familienbesuch, Familienleben, Familienmitglied, Familienmitglieder, Geschwister, Großeltern, Großeltern, Großmutter, Hellmuth, Hildes Eltern, Kin, Mutter, Oma, Onkel, Pate, Rolands Eltern, Siegfried, Vater, Verwandte	Familie
Abendmahlfeier, Advent, Adventgottesdienst, Adventszeit, Begräbnis, Erntedankfest, Erster Advent, Feier, Feierlichkeiten, Feiern, Ferien, Fest, Festlichkeiten, Geburstagsgeschenk, Heiligabend, Heilige drei Könige, Hochzeit, Hochzeitsgeschenk, Hochzeitstag, Jahrmarkt, Jubiläum, Kirmes, Konfirmation, Kriegsweihnacht, Lazarettbescherung, Muttertag, Ostern, Pfingsten, Polterabend, Rauhnacht, Silvester, Silvesterfeier, Silvesternacht, Sylvester, Tag der Wehrmacht, Taufe, Trauer, Verlobungstag, Volksfest, Weihnachten, Weihnachtessen, Weihnachtsbaum, Weihnachtsessen, Weihnachtsfeier, Weihnachtsfeiertage, Weihnachtsfest, Weihnachtsgeschenk, Weihnachtsgeschenke, Weihnachtslieder, Weihnachtsmann, Weihnachtsmesse, Weihnachtsvorbereitungen, Weihnachtszeit	Feste
Film, Filme, Französischer Film, Frontkino, Heinz Rühmann, Kino, Kinoabend, Kinofilm, Liebesfilm, Propagandafilm, Schauspieler, Wochenschau, Zarah Leander	Film
Bild, Bilder, Foto, Fotoabzüge, Fotoalben, Fotoapparat, Fotofilm, Fotograf, Fotografieren, Fotos, Fotostudio, Schnappschuss	Fotografie
Ablenkung, Amüsieren, Ausgang, Ausgehen, Auszeit, Freizeit, Sonnenbaden, Spielen, Spielzeug, Unterhaltung	Freizeit
Freund, Freunde, Freunden, Freundschaft, Freundschaftsvertrag, Kränzchen, Männerfreundschaft	Freunde
Adolf Hitler, Führer, Führergeburtstag, Führerprinzip, Hitler	Führer
Achtung, Alleinsein, Angst, Antipathie, Aufregung, Autonomie, Begehrlichkeit, Behagen, Beherrschung, Behinderung, Bewunderung, Bitterkeit, Dankbarkeit, Demut, Demütigung, Depression, Dienstbereit, Durchhalten, Eifersucht, Einfühlung, Einsamkeit, Einssein, Enttäuschung, Erfolg, Erfüllung, Erlebnis, erleichtert, Erleichterung, Erregung, Ferne, Freude, Geborgenheit, Gefahr, Gefühle, Gemeinschaft, Gemüt, Genuß, Glücksgefühl, Heimweh, Hoffnung, Innerer Frieden, Intimität, Kriegsmüdigkeit, Kriegsmüdigkeit, Kriegsmüdigkeit, Langeweile, Launenhaftigkeit, Leben allein, Leid, Leidenschaft, Liebe, Liebesbekenntnis, Liebesbekundungen, Melancholie, Mitleid, Motivation, Müdigkeit, Muße, Mut, Nähe, Neid, Neugier, Optimismus, Pflichtgefühl, Probe, Prüfung, Resignation, resignieren, Respekt, Scham, Schuld, Schwermütigkeit, Sehnsucht, Sehnsüchte, Sicherheit, Sinn, Sinnlichkeit, Sorge, Sorgen, Stärke, Stimmung, Stolz, Subjektivität, Tränen, Traum, Träume, Traurigkeit, Treue, Trost, Trübsal, Trübsinn, Ungeduld, Ungewissheit, Unruhe, Unsicherheit, Unstimmigkeit, Unzufriedenheit, Veränderung, Verbundenheit, Verdorbenheit, Vergewisserung, Versuchung, Vertrauen, Vertrautheit, Verzicht, Verzweiflung, Warten, Weinen, Weltschmerz, Wichtigkeit, Zufriedenheit, Zuneigung, Zurückweisung, Zusicherung, Zuversicht, Zwang, Zweifel, Zynismus	Gefühle

Old Keywords	New Keyword
Bankkonto, Besitztum, Darlehen, Erbe, Finanzen, Gehalt, Geld, Geldschulden, Geldsendung, Haushaltskosten, Kostgeld, Krankenkasse, Lebensversicherung, Miete, Mietvertrag, Pension, Rechnung, Rechnungen, Reichsbesondung, Reichtum, Sold, Sparen, Sparprogramm, Spenden, Spendensammlung, Steuer, Vermietung, Versicherung, Währung	Geld
Geburtstagsüberraschung, Geburtstagsgeschenk, Blumengeschenk, Geschenk, Geschenke, Mitbringsel, schenken	Geschenke
Ehefrau, Frauen, Frauen der Kameraden, Frauen im Nationalsozialismus, Frauenarbeit, Frauenbild, gender, Geschlecht, Geschlechter, GeschlechterBeziehungen, GeschlechterBeziehungenen, Geschlechterrolle, Geschlechterrollen, Hausfrau, Jungfräulichkeit, Männlichkeit, Mütterlichkeit, Mutterliebe, Mutterrolle, Mutterschaft, Vaterrolle	Geschlechterrollen
Anschlag, Artzneimittel, Arzt, Arzt/Patient, Arztbesuch, Attest, Behandlung, Doktor, Erkältung, Fehlgeburt, Frauenarzt, Gebärfähigkeit, Geburtsfehler, Genesung, Geschlechtskrankheiten, Gesundheit, Gesundheitsverhältnisse, krank, Krankeit, Kranken, Krankenpflege, Krankenschwester, Krankheit, Krankheiten, Läuse, Magenschmerzen, Medizin, Menstruation, Menstruationskalender, Operation, Parasiten, Pflege, Prothese, Röntgenaufnahmen, Schlafstörung, Schmerz, Typhusangst, Untersuchung, Verpflegung, Zahnarzt, Zahnarztbesuch, Zähne, Zahnschmerzen	Gesundheit
Aberglaube, Astrologie, Atheismus, Bibel, Bibelvers, Bibelzitat, Christentum, Deutsche Christen, Deutsche Glaubensbewegung, Gebet, Glaube, Glauben, Glück, Gott, Gott, Gottvertrauen, Heiland, Jesus, Konfession, Mythos, Numerologie, Protestantismus, Religion, Säkulärismus, Schicksal, Seele, Sternbilder, Sternsänger, Sünde, Unglück	Glaube
Barbarei, Eroberung, Eroberungskrieg, Flüchtlinge, Generalplan Ost, Hungerplan, Kriegsschäden, Massenmord, Opfer	Gräueltaten
Kochen, Backen, Basteln, Besorgungen, Einkochen, Handarbeit, Hausarbeit, Haushalt, Hauswirtschaft, Näharbeit, Nähen, Nähmaschine, Ordnung, Putzen, Saubermachen, Stickarbeit, Strickarbeit, Stricken	Hausarbeit
Ausstattung, Beleuchtung, Bettwäsche, Einrichtung, Elektrogeräte, Hausstand, Heizung, Herrenzimmer, Kessel, Kohle, Laterne, Möbel, Möbel für Herrenzimmer, Möbelkauf, Möbellieferung, Ofen, Taschenlampe, Wärmefläsche, Wasserleitung	Hausrat
Heimat, Heimatfront	Heimat
Bad, Baden, Dusche, duschen, Friseur, Frisur, Hygiene, Hygieneartikel, Hygieneprodukte, Reinigen, Reinigung, Waesche, Wäsche, Waschen, Wäschewaschen	Hygiene
Gefährte, Kamerad, Kameraden, Kameraden, Kameradschaft	Kameraden
Adoption, Baby, Einzelkind, Kind, Kinder, Kinderaufführung, Kinderbetreuung, Kindererziehung, Kinderfürsorge, Kinderkriegen, Kinderwunsch, Kindheit, Nachwuchs	Kinder
Chor, Gottesdienst, Kantorei, Kirche, Kirchenbesuch, Kirchenbräuche, Kirchenbücher, Kirchenchor, Kleriker, Kloster, Pastor, Predigt, Singstunde, Singstunden	Kirche
Geheimnis, Geheimnisse, Gerüchte, Gespräch, Geständnis, Gruss, Grüße, Klatsch, Kommunikation, Kommunikationsmittel, Kosename, Kosenamen, Missverständnis, Salutschießen, Schelte, Selbstzensur, Streit, Streitigkeit, Telefon, Telefongespräch, Tratsch, Willkommen, Witz, Witzemachen	Kommunikation

Old Keywords	New Keyword
Aussehen, Bart, Blut, Geruch, Gewicht, Haare, Klimakterium, Koerper, Körper, Körperbild, Körperlichkeit, Körperpflege, Leibe, Muskelkater, Nacktheit, Weisheitszahn	Körper
Kriegsgemeinschaft, Kriegsgesellschaft, Volk, Volksgemeinschaft	Kriegs- /Volksgemeinschaft
Gasvergiftung, Gefallene, Gefallener, Heimatlosen, Kampfspuren, Kriegsgefangene, Kriegsspur, Kriegstote, Notzeit, Schussverletzung, Tote, Vergiftung, Verletzung, Verlustlisten, Vertreibung, Verwundete, Verwundung, Zerstörung	Kriegsfolgen
Afrikafeldzug, Afrikafront, Balkanfeldzug, Fall Blau, Feldstellung, Feldzug, Front, Fronterfahrung, Frontleben, Leningrader Blockade, Operation Torch, Osteinsatz, Ostfront, Ostkrieg, Pazifikkrieg, Russlandfeldzug, Schlacht um Moskau, Westfront	Kriegsschauplatz
Abwehrmassnahmen, Angriff, Anleitung, Besatzung, Besatzungsregeln, Einmarsch, Einsatz, Endsieg, Entwaffnung, Evakuierung, Frieden, Kampf, Kapitulation, Krieg, Kriegsalltag, Kriegsbeginn, Kriegsdauer, Kriegsende, Kriegserfolg, Kriegserklärung, Kriegsgeschehen, Kriegsmaschine, Kriegsverlauf, Kriegsvorbereitung, Panzerschlacht, Partisanen, Schießerei, Seeschlacht, Sieg, Stationierung, Waffenstillstand	Kriegsverlauf
Brauchtum, Kultur, Kulturveranstaltung, Mode, Modern, Modernisierung, Modernität, Modetorheit, Tradition	Kultur
Amerikaner, Ausland, Ausländer, Buren, deutsch-indische Kinder, Englaender, Engländer, Fremde, Fremde Völker, Fremdenfeindlichkeit, Fremdwahrnehmung, Griechen, Griechinnen, Indianer, Kroaten, Moschee, Russen, Sinti und Roma, Stereotyp	Kulturkontakt
Ausstellung, Kunst, Kunstgalerie, Künstler, Maler, Malerei, Rembrandt, Scherenschnitt, Zeichnung	Kunst
Landschaft, Meer, Region, Strand, Welt	Landschaften
Geburt, Laufbahn, Schwangerschaft, Sterben, Tod, Tot, Leben	Lebenszyklus
Heimatgeschichten, Alice im Wunderland, Autor, Buch, Buchausschnitt, Bücher, Buchhandlung, Buchzitate, Dichter, Dorfbücher, Gedicht, Gedichte, Georg von Lübeck, Hänsel und Gretel, Heldendichtung, Lesen, Literatur, Märchen, Märchenvorstellung, Poesie, Roman, Sagen, Schriftsteller, Tornisterschrift, Wilhelm Schäfer, Wilhelm Stapel, Woerterbuch	Literatur
Alarm, Bombardierung, Entwarnungssirenen, Fliegeralarm, Fliegerangriffe, Luftalarm, Luftangriff, Luftangriffe, Luftkrieg, Luftschutz, Luftunfall, Luftwaffe, Verdunkelung	Luftkrieg
Adolf Rosenberg, Bismarck, Chamberlain, Churchill, Diktatur, Diplomaten, Erwin Rommel, Francois Darlan, Göbbels, Göring, Hermann Göring, Ion Antonescu, Joseph Goebbels, König von Rumänien, Mussolini, Regierung, Regime, Rommel, Roosevelt, Rosenberg, Rudolf Hess, Staatsmann, Stalin	Machthaber
Auszeichnung, Dienstorden, Eid, Etappe, Feldleben, Helferinnen, Kriegseinsatz, Lageplan, Lager, Militär, Militärdienst, Militarismus, Militärübung, Militärzeit, Nachrichtenhelferin, Orden, Soldatenalltag, Soldatenleben, Truppenbetreuung, Uniform, Verteidigung, Wehrdienst, Wehrmacht, Weltkrieg	Militär
Abreise, Abschied, Ankunft, Aufenthalt, Auswandern, Auto, Bahn, Bahnhof, Boot, Bootfahrt, Bus, Durchreise, Eisenbahn, Fahrplan, Fahrrad, Fahrt, Fahrzeug, Gepäck, Heimkehr, Heimreise, Karte, Marschbefehl, Reise, Reisen, Strassenbahn, Transport, Truppenbewegung, Truppentransport, Truppentransporte, umziehen, Umzug, Verabschiedung, Verkehr, Verkehrsmittel, Verlegung, Versetzung, Zug, Zugfahrt	Mobilität

Old Keywords	New Keyword
Veranstaltung, Deutscher Jazz, Franz Schubert, Hausmusik, Jazz, Kantate, Klassische Musik, Klavier, Klavierspielen, Komponist, Konzert, Lied, Lieder, Mozart, Musik, Musizieren, Orgel, Orgelmusik, Philharmonie, Schallplatte, Schlager, Singen, Soldatenlieder, Volkslied	Musik
Arisierung, Aufklärungsbücher, Deutschtum, Existenzkampf, Fahne, Fanatismus, Gleichschaltung, Hakenkreuz, Hakenkreuzfahne, Herrenvolk, Ideologie, Konformität, Lebensraum, Nationalhymne, Nationalismus, Nationalsozialismus, nationalsozialistische Ideologie, NS-Ideologie, Opferbereitschaft, Ostsiedlung, OT, Parole, Propaganda, Schicksalsgemeinschaft, Siedlungsgemeinschaft, Siegeswille, Siegeswillen, Totaler Krieg, Vernichtung, Vernichtungskrieg, Volksdeutsche, Volkstumspolitik, Weltanschauung, Zerstörungskrieg	Nationalsozialismus
Baum, Blumen, Haustier, Insekten, Lärmverschmutzung, Licht, Lichtverschmutzung, Mond, Natur, Naturverehrung, Pflanzen, Sonne, Sterne, Temperatur, Tiere, Tierschutz, Trockenblume, Ungeziefer, Vögel, Wanzen, Wärme, Wasser, Wesen, Zucht	Natur
Diensträume, Dienststelle, Einquartierung, Flugzeugplatz, Gastbetrieb, Gaststätte, Hafen, Hotel, Judenfriedhof, Kaserne, Kasernen, Kasernenleben, Kneipe, Krankenhaus, Lazarett, Marinehafen, Markt, Soldatenheim, Strassen, Unterkunft, Waschhaus, Wehrmachtheim, Zimmerbelegung	Öffentliche Räume
Oper, Operette, Richard Wagner	Oper
Abstand, Auseinandersetzung, Beziehungen, Brautwerbung, Distanz, Ehering, Entfernung, erste Liebe, FernBeziehungen, Freundin, Heirat, Kennenlernen, LiebesBeziehungen, LiebesBeziehungenen, Liebeserklärung, Liebessehnsucht, Liebschaften, Mord, PaarBeziehungen, Paarsein, Rat, Ratschlag, Selbstmord, Trennung, Treueversprechen, Verlangen, Verlobung	Paarbeziehung
Arbeitsdienst, BDM, Deutsches Frauenwerk, Deutsches-Rotes-Kreuz, Frauendienst, Frauendienst, Frauenschaft, Frauenwerk, Hitlerjugend, HJ, KdF, nationalsozialistische Organisationen, NS Volkswohlfahrt, NS-Frauenschaft, NS-Massenorganisation, NSDAP, NSV, Partei, Parteimitgliedschaft, Parteipolitik, Reichsarbeitsdienst, Reichsparteitag, Rotes Kreuz, Rotes Kreuz, Rotkreuzschwester, SS, Volkswohlfahrt, WHW, Winterhilfswerk	Partei
Ansprache, Attentat, Aussenpolitik, Denunziation, Eiserne Garde, Freimaurerei, Kinderlandverschickung, Kritik, Neuordnung, Nichtkonformität, Nonkonformität, Normalität, persönliche Rechte, Politik, politische Befugnisse, Regimekritik, Reichstag, Siedlungsplan, Siedlungspläne, Tagespolitik, Umsiedeln, Umsiedlung, Vaterland, Vaterlandsdienst, Vaterlandsliebe, Wahl, Weltpolitik, Widerstand	Politik
Anstarren, Betrug, Ehrgeiz, Eigensinn, Eigenständigkeit, Eigenverantwortung, Empathie, Entscheidung, Entschluss, Entschuldigen, Ermahnung, Gastfreundschaft, Geselligkeit, Gleichgültigkeit, Hektik, Hilfe, Masken, Mitmachen, Routine, Schikane, Schlafen, Schwindel, Selbstbeobachtung, Selbstdarstellung, Selbstreflexion, Selbstverständnis, Selbstvertrauen, Verhalten	Praktiken
Badezimmer, Eigenheim, Elternhaus, Haus, Hause, Heim, Küche, Privat, Privatbereich, Privatleben, Privatsphäre, Schlafzimmer, Stube, Toilette, Wohnen, Wohnung, Wohnzimmer, Zuhause	Private Räume
Ahnenforschung, Erbkrankheit, Euthanasie, Hautfarbe, Kulturgefälle, Menschensarten, Rasse, Rassenbiologie, Rassenkunde, Rassenpolitik, Rassismus, Sozialdarwinismus, Wesensverwandtschaft	Rassismus
Führerrede, Hitlerrede, Rede, Vortrag, Vortragsreihe	Reden
Nachrichten, Radio, Radioprogramme, Rundfunk, Wunschkonzert	Rundfunk

Old Keywords	New Keyword
Adresse, Brief, Brief an Eltern, Brief von Bekannten, Briefe, Briefeschreiben, Briefform, Briefgeheimnis, Briefmarke, Briefpapier, Briefschaftenkiste, Briefschulden, Briefumschlag, Briefverlust, Briefwechsel, Buchstabieren, Feldpost, Geheimschrift, Handschrift, Kartons, Kriegspost, Kriegszensur, Liebesbrief, Luftpost, Päckchen, Paket, Pakete, Papier, Post, Postamt, Postbote, Postdaten, Postdauer, Postkarte, Postsperre, Postüberwachung, Postverkehr, Postverkehr, Schreiben, Schreibgerät, Schreibgeräte, Schreibmaschine, Schreibpapier, Schrift, Stempel, Stenographie, Stift, Stifte, Telegram, Telegramm, Text, Tinte, Zensur, Zitat aus dem Brief	Schreiben
Attraktivität, Aussereheliche Beziehungenen, Erotik, Fensterln, Flirt, Fremdgehen, Fruchtbarkeit, Geschlechtsverkehr, Kondom, Kuss, Küsse, Küssen, Lust, Prävention, Prostitution, Sex, Sexualität, Sexualnot, Sexualwerben, sexuelle Belästigung, sexuelle Normen, Sittlichkeit, Unzucht, Zärtlichkeit, Zärtlichkeiten	Sexualität
Bergsteigen, Schlittenfahrt, Wandern, Wanderung, Gymnastik, Schwimmen, Skifahren, Sport, Tango, Tanz, Tanzen, Tanzstunde, Tanzverbot, Turnen, Volkstanz	Sport/Tanz
Dialekt, Duzen, Fremdsprachen, Metapher, Militärische Sprache, Muttersprache, Soldatensprache, Sprache, Sprachveränderungen, Volksmund	Sprache
Ebenbürtig, Ehrbarkeit, Niveau, Präsentation, Skandal, Stand, Standesunterschied, Unterordnung	Status
Fallschirmstaffel, Fallschirmtruppen, Flieger, Flugschau, Infanterie, Kriegsmarine, Marine, Matrosen, Offizier, Offiziere, Piloten, Rekruten, Soldat, Soldaten, Truppe, Truppendisziplin, Wehrmachthelferin, Wehrmachtshelferin, Zivilverteidigung	Streitkräfte
Aufführung, Fronttheater, Schauspiel, Theater, Theaterbesuch, Theaterstück, Volksstück	Theater
Ausflug, Ausflüge, Spazieren, Spaziergang, Spaziergang, Andenken, Ansichtskarte, Besichtigung, Erinnerungsfotos, Fotoandenken, Fotoerinnerung, Krieg als Reise, Sehenswürdigkeiten, Souvenir, Stadtbesichtigung, Tourismus, Urlaubsfotos, Villa	Tourismus
Ernteurlaub, Fronturlaub, Heimaturlaub, Kururlaub, Rückkehr, Sommerfrische, Urlaub, Urlaubsende, Urlaubserlaubnis, Urlaubssperre	Urlaub
Antike, Beziehungensgeschichte, Entstehungsgeschichte, Erinnerung, Erinnerung an Hochzeit, Erinnerung an Kennenlernen, Erinnerung an Verlobung, Erinnerungen, Erster Weltkrieg, Geschichte, Kulturgeschichte, Leibesgeschichte, Liebesgeschichte, Vergangen, Vergangenheit, Weltgeschichte, Zäsuren	Vergangenheit
Ausrüstung, Flak, Flugzeug, Flugzeuge, Gas, Gasmasken, Gift, Lützow, Panzer, Schiff, Sprengungen, U-Boote, Waffe, Waffen	Waffen
Anstand, Aufopferung, Charakter, Defätismus, Deutschsein, Ehre, Ehrlichkeit, Einzelgänger, Eitelkeit, Feigheit, Fürsorge, Geduld, Gerechtigkeit, Geschmack, Gewissen, Gleichberechtigung, Gleichheit, Haltung, Hass, Häuslichkeit, Heldentod, Heldentum, Herrschaft, Hierarchie, Ideale, Individualität, Kampfbereitschaft, Kampfeswille, Keuschheit, Lebensführung, Lebenswille, Menschen, Missachtung, Misstrauen, Moral, Nächstenliebe, Nützlichsein, Patriotismus, Persönlichkeit, Pflicht, Realist, Reinheit, romantische Liebe, Rücksicht, Sauberkeit, Schlichtheit, Schönheit, Schöpfergeist, Selbstbeherrschung, Selbstlosigkeit, Selbstständigkeit, Selbstüberwindung, Sitten, Sittlichkeitsnorm, Soziale Ungerechtigkeit, Tapferkeit, Temperament, Traulichkeit, Tugend, Unabhängigkeit, Unehelichkeit, Unerwiderte Liebe, Unordnung, Verantwortung, Vorurteile, Vorwürfe, Wahrheit, Wert	Werte

Old Keywords	New Keyword
Frost, Hochwasser, Kälte, Klima, Schnee, Schneesturm, Wetter, Winterwetter	Wetter
Ausgaben, Arbeitseinsatz, Arbeitshilfe, Arbeitskräftemangel, arm, Armut, Beschlagnahmung, Beute, Bevorratung, Einkauf, Einkäufe, Einkaufen, Eintopf, Entbehrung, Ernte, Ersatzarbeitskräfte, Ersatzlebensmittel, Ersatzlehrer, Geschäft, Hamstern, Hungersnot, Importwaren, Inflation, Kaufen, Kleiderkarten, Kleidermarken, Knappheit, Konsum, Konsumbeschränkungen, Konsumgesellschaft, Konsumwünsche, Kriegsarmut, Kriegsbeute, Kriegswirtschaft, Landwirtschaft, Lebensmittel, lebensmittelbeschaffung, Lebensmittelkarten, Lebensmittelmarken, Mangel, Materialknappheit, Nahrungsmangel, Organisieren, Preise, Ration, Rationierung, Ressourcen, Rüstungsindustrie, Schleichhandel, Schwarzmarkt, schwere Zeit, Stoffmangel, Stromausfall, Tauschhandel, Versorgung, Versorgungsschwierigkeiten, Vorräte, Warenangebot, Wirtschaften, Wohlfahrt, Wohlstand, Wohnungsmarkt, Wohnungssuche, Wucher, Zahlkarte, Zwangsarbeit, Zwangsarbeiter	Wirtschaft
Astronomie, Ethnologie, Geografie, Geographie, Germanistik, Logik, Philosoph, Philosophie, Psychologie, Rechnen, Religionswissenschaft, Theologie	Wissenschaft
Abend, aufstehen, Datum, Feierabend, Feiertag, Feiertage, Frühling, Frühlingsbeginn, Geburtstag, Herbst, Jahresende, JahresKalender, Jahrestag, Jahreszeit, Johannistag, Kalendar, Kalender, KriegsKalender, Kriegszeit, Monat, Neujahr, Sonntag, Tagesablauf, Tagesgeschehen, Uhrzeit, Wecken, Winter, Wochenende, Zeitumstellung, Zeitverschiebung, Zeitzone, Zyklus	Zeit
Artikelausschnitt, Bekanntmachung, Illustrierte, Medien, NS Medien, Presse, Sondermeldung, Verlag, Wehrmachtbericht, Zeitschrift, Zeitung	Zeitungen
Freiheit, Hoffnung auf Kriegsende, Plan, Reiseplan, Reisepläne, Vorbereitung, Vorbereitungen, Vorfreude, Vorstellung, Wiedersehen, Wünsche, Ziel, Zukunft, Zukunftspläne, Zukunftsplanung, Zukunftsträume	Zukunft
Alltag, Bauen, Blatt, Dorf, Feder, Gebäude, Gemeinde, Gesellschaft, Gesellschaftsordnung, Großstadt, Großstadt, Güter, Holz, Klempner, Michael I., Name, Periode, Personallage, Programm, Rationalisierung, Reich, Renovierung, Schaden, Schwierigkeiten, Selbst und Gemeinschaft, Stadt, Turm, Umbau, Umgebung, Unfall, Versäumnis, Verspätung, Verzögerung	deleted

Place names are not part of the thematic keywords in AiK anymore. Instead, the letters are separately tagged with the places mentioned in them and the locations of sender and recipient. The keywords “Empfohlen von FAG” (“Recommended by Freie Altenarbeit Göttingen”) and “Unvollständig” (“Incomplete”) are not thematic and are therefore not considered in the thesis.

A.2 Glossary of Keywords

All keyword definitions were written by Laura Fahnenbruck (Fahnenbruck et al 2023). Currently, they are only available for registered members on the *Alltag im Krieg* website.

Allianzen: Alliierte, Achsenmächte, sowie die in Bezug auf den Zweiten Weltkrieg neutralen Länder.

Antisemitismus: Gefühle, Haltungen und Maßnahmen gegen Juden und Jüdinnen gerichtet.

Alter: Umfasst auch die Themen, Generationen, Generationenunterschiede und Lebensaltersabschnitte.

Arbeit: Umfasst alle Arten von Arbeit und Dienst, z.B. militärischer Dienst, nicht aber Hausarbeit und Ausbildung.

Aus-/Bildung: Umfasst Ausbildung, Bildung, Erziehung, Lehren, Lernen und Übung, z.B. auch militärisches Exerzieren.

Ausweise: Persönliche Dokumente wie Ausweise, Stammbäume, Urkunden sowie das Nachforschen und Nachweisen hiervon.

Baukunst: Umfasst Architektur, Bausubstanz und damit verbundene Menschen und Tätigkeiten, z.B. Baumeister, Stadtplanung.

Begegnungen: Alle Arten von informellen zwischenmenschlichen Treffen und die Beschreibungen davon.

Behörde: Offizielle Verwaltungsinstanzen inklusive Polizei und Justiz.

Bekannte: Umfasst auch Kollegen, Kolleginnen, Nachbarinnen und Nachbarn, nicht aber Familie, Freunde und Kameraden.

Bekleidung: Umfasst Kleidung aller Art, inklusive Hüte, Schuhe, Schmuck und Schminke.

Bevölkerung: Umfasst den geografischen und soziologischen Begriff von Menschenmassen und Demografie, nicht aber die konkrete Beschreibung von Begegnungen mit Menschen (siehe Kulturkontakt), und nicht die politische Begrifflichkeit (siehe Kriegs-/Volksgemeinschaft, Rassismus).

Essen/Trinken: Umfasst auch Nahrung, Ernährung, Alkohol und Rauchwaren sowie die Tätigkeiten Essen, Trinken, Rauchen.

Familie: Umfasst den näheren Familienkreis der Schreibenden, also Geschwister, Schwager und Schwägerinnen, Eltern und Kinder, sowie die weiteren Verwandten, umfasst auch Familienbesuch.

Feste: Umfasst Festlichkeiten und Feiern aller Art, nicht aber das bloße Aufzählen oder Erzählen der Feiertage (siehe Zeit).

Film: Umfasst Filme, Filmbesuche, Kino, Schauspieler und Schauspielerinnen, sowie Wochenschau.

Fotografie: Umfasst Fotoapparate, Fotografien, Fotoalben, das Fotografieren, Fotografen, Fotoentwicklung und Fotostudios, nicht aber die explizite Erinnerungsfunktion, wenn von Fotoandenken und Erinnerungsfotos die Rede ist (siehe Tourismus).

Freizeit: Die freie Zeit außerhalb von Arbeit und Dienst, sowie Freizeitbeschäftigungen und Veranstaltungsbesuche, sofern es bei diesen nicht konkret um ein Thema geht (siehe Film, Sport/Tanz, Theater, Literatur).

Freunde: Nähere, nicht-verwandte Beziehungen, nicht aber Kameraden oder Bekannte.

Führer: Umfasst Adolf Hitler als Person oder Politiker.

Gefühle: Umfasst den Austausch von emotionalen Erfahrungen, nicht aber die Bewertung von Ereignissen oder die Thematisierung von Normen (siehe Werte).

Geld: Umfasst auch die konkrete Benennung von Wertmarken, Tauschhandel, Sold, Ausgaben und Überweisungen, nicht aber die allgemeine ökonomische Lage (siehe Wirtschaft).

Geschenke: Umfasst auch im Brief mitversendete Geschenke und Mitbringsel.

Geschlechterrollen: Umfasst die Thematisierung von Erwartungen und Normen bezüglich von Geschlechtern, Männlichkeit, Weiblichkeit, Ehepartnern, Ehepartnerinnen, Väterlichkeit und Mütterlichkeit, nicht aber Sexualnormen (siehe Sexualität).

Gesundheit: Umfasst Gesundheit und Krankheit, Arztbesuche, Krankenhaus, Lazarett, Verletzung, (In-)Validität, nicht aber Tod, Geburt (siehe Lebenszyklus) und Schwangerschaft, Menstruation (siehe Körper) und Hygiene (siehe Hygiene) und nicht Appelle an die 'Volksgesundheit' (siehe Nationalsozialismus).

Glaube: Umfasst alle Glaubensphänomene und die Thematisierung von Zuversicht an eine höhere Existenz, wie Gott, Glaube, Aberglauben, Säkularismus und alle Religionen, nicht aber die konkrete Auslebung von Religion in der Kirche oder die Thematisierung von Kirchenpersonen und Kirche als Institution und Ort (siehe Kirche).

Gräueltaten: Umfasst auch Aushungerung, Eroberung, (Massen-)Erschießung, Massenmord, Zwangsumsiedlung und sonstige Verbrechen gegen die Menschlichkeit, systematisch oder inzidentell, nicht aber allgemeinere Kriegsfolgen wie Flucht, Bombardierung, Zwangsumsiedlung (siehe Kriegsfolgen) oder politische Pläne zu Gräueltaten wie Generalplan Ost, Umsiedlung, Konzentrationslager, Vernichtungskrieg, Gaskammern (siehe Nationalsozialismus) und nicht Vergewaltigungen, die nicht im Kriegsgebiet stattfanden (siehe Sexualität).

Hausarbeit: Umfasst Hauswirtschaft, Besorgungen und alle Tätigkeiten, die mit Haus und Garten verbunden sind.

Hausrat: Einrichtungs- und Gebrauchsgegenstände in Haus und Garten, wie Ofen, Möbel, Lampen, Service, sowie die Hochzeitsausstattung.

Heimat: Umfasst auch Heimatgefühl, Heimatmuseum, Heimatlosigkeit und Heimatfront, wenn der Rückhalt der deutschen Bevölkerung oder einer konkreten Person für die Kriegsschauplätze im metaphorischen Sinn angesprochen ist, nicht aber wenn dieser Rückhalt durch konkretes Handeln angesprochen wird (siehe Kriegs-/Volksgemeinschaft).

Hygiene: Umfasst auch Körperpflege als Tätigkeit und diesbezügliche Gegenstände wie Seife, Bad und Sauberkeit als Idee, nicht aber 'Rassenhygiene' (siehe Rassismus) und gesundheitliche Aspekte (siehe Gesundheit).

Kameraden: Umfasst Kameraden und Kameradinnen im Militär und in den zivilgesellschaftlichen Organisationen der NSDAP, Kameradschaft, Kameradschaftsehe, Kameradschaftsgefühle, Appelle an Kameradschaftlichkeit, nicht aber zwischenmenschliche Beziehungen mit Familie, Bekannten oder Freunden.

Kinder: Umfasst auch Kinderfrage, Kinderwunsch, Kinderplanung, Kindererziehung, Kindheit, Nachwuchs, nicht aber Kinderschar und Schar (siehe Arbeit) oder die allgemeinere Erziehung zum erwachsenen, ausgebildeten Menschen (siehe Aus-/Bildung).

Kirche: Umfasst Kirche als Institution und Ort, sowie Tätigkeiten und Menschen, die damit verbunden sind, zum Beispiel Pfarrer, Chor, Singestunde, Bibel, Predigt, nicht aber abstraktere Ideen zur Kirche oder die Auseinandersetzung mit Religionen und Glaubensfragen (siehe Glaube).

Kommunikation: Umfasst zwischenmenschliche Kommunikation in Gesprächen und am Telefon, sowie Kommunikationsmodi wie Geheimnisse, Klatsch, Streitigkeiten, Koseworte, Witze, Gerüchte, Lästern, nicht aber die schriftliche Kommunikation und postalischer Verkehr (siehe Schreiben).

Körper: Umfasst alle Benennungen von Körperlichkeit oder Körperteilen sowie ihren Eigenschaften und Funktionen, wie Gebrechlichkeit, Schmerzen, Frieren, Geruch, Füße, Körpertemperatur, Menstruation, Schwangerschaft, nicht aber Körperhygiene (siehe Hygiene) und Körpergesundheit (siehe Gesundheit).

Kriegs-/Volksgemeinschaft: Umfasst Handlungen und Ideen, die den Zusammenhalt von den nicht-jüdischen Deutschen, die in die deutsche Mehrheitsgesellschaft inkludiert waren, erwirken, beschwören oder bestätigen sollten, sei es an der Front, in den besetzten Gebieten oder im gesamten Deutschen Reich, bzw. in der Gesellschaft derjenigen, die in die Volksgemeinschaft inkludiert werden sollten. Umfasst nicht diejenigen, die ausgeschlossen wurden (siehe Antisemitismus, Rassismus).

Kriegsfolgen: Umfasst die konkreten Folgen des Krieges, an den Kriegsschauplätzen und in der Heimatfront, wie Flucht, Kriegsschäden, Zerstörung, Kriegstote, Gefallene, Ausbombardierte, Aufnahme von Flüchtlingen, Zwangsumsiedlung, Armut und Hunger als Folge von Kriegsstrategie.

Kriegsschauplatz: Frontgebiete, Schlachten, Schauplätze von Luft- und Seeschlachten, nicht aber Heimatfront, besetzte Gebiete und Etappenstellungen.

Kriegsverlauf: Kriegsereignisse im Allgemein sowie Beschreibungen konkreter Ereignisse und das Thematisieren von Sieg, Niederlage, Kriegsende, Kriegsverantwortung, Kriegsstrategien.

Kultur: Umfasst einen breiten Kulturbegriff, also menschliche Ideen, Handlungen, und Produkte von Kultur im Allgemeinen, wie Brauchtum, Mode, Modernität, Modernisierung, Tradition, Sitten, nicht aber Kultur im engeren Sinne (siehe Film, Literatur, Theater, Oper) und nicht kulturelle Begegnungen, Fremdwahrnehmung und kulturelle Stereotype (siehe Kulturkontakt).

Kulturkontakt: Umfasst konkrete kulturelle Begegnungen und Begegnungen mit 'Anderen', an der Front, in den besetzten Gebieten und in Deutschland, sowie das Denken über 'Kulturen', Fremd- und Eigenwahrnehmung und Stereotype, nicht aber das konkrete Thematisieren von nationalsozialistischen Varianten davon, wie 'Wesensarten', 'Rassenhygiene', 'Volkstum' (siehe Antisemitismus, Rassismus, Nationalsozialismus).

Kunst: Umfasst Malerei, Zeichnungen, und andere bildende Künste, nicht aber Gebrauchsfotografien (siehe Fotografie) und Darstellung (siehe Film, Theater).

Landschaften: Umfasst Landschaftsbeschreibungen, nicht aber Naturbeschreibungen (siehe Natur) und (Groß-)Städte (siehe Öffentliche Räume).

Lebenszyklus: Lebensabschnitte und Initiationsriten im Laufe eines Lebens, wie Geburt, Tod, Volljährigkeit, Mutterwerdung, nicht aber die Feierlichkeiten zu einem solchen Anlass selbst, wie Hochzeit, Schulanfang (siehe Feste).

Literatur: Umfasst auch Autoren, Bücher, Dichter, Denker, Gedichte.

Luftkrieg: Umfasst konkrete Bestandteile des Luftkrieges, wie Angriffsflüge, Luftalarm, Flugzeugtypen, Bombardierung, Bombenalarm, Verdunkelung.

Machthaber: Politische Führungspositionen und Politiker oder Gruppen von Machthabenden in Machtpositionen, nicht aber Adolf Hitler (siehe Führer) und nicht Vorgesetzte und Chefs (siehe Behörden, Arbeit).

Militär: Umfasst das Militärische als Idee und als Alltagspraxis, wie die Dienstzeit im Militär, militärische Orden, Etappendasein, Uniformen, Militärlager, Soldatenalltag, Truppenbetreuung, militärische Aufgaben, nicht aber die Einübung militärischer Handlungen und Tugenden (siehe Arbeit, Aus-/Bildung, Kameraden), und nicht die allgemeinere militärische Strategie, Kriegsschauplätze oder den Verlauf des Krieges (siehe Politik, Nationalsozialismus, Kriegsschauplatz, Kriegsverlauf), sowie nicht konkrete militärische Verstöße (siehe Gräueltaten).

Mobilität: Umfasst alle Fortbewegungsweisen, Verkehrsmittel und Ortswechsel, wie Züge, Zugfahrt, Truppenbewegung, Truppenverlegung.

Musik: Umfasst auch Musiker, Noten, Dirigenten, Konzerte, Sängerinnen, nicht aber Singestunde und Kirchenchor (siehe Kirche), Wunschkonzert (siehe Rundfunk) und Oper (siehe Oper).

Nationalsozialismus: Umfasst die nationalsozialistische Ideologie und dazugehörige Taten oder Pläne, Begriffe und Konzepte, wie Schicksalsgemeinschaft, Hakenkreuz, Hakenkreuzfahne, Umsiedlung, Konzentrationslager, Vernichtungskrieg, Gaskammern, Propaganda, Lebensraum, Existenzkampf, Deutschtum, Opferbereitschaft, Wesensverwandtschaft, Volkstum, nicht aber die Praktiken und Umsetzung (siehe Rassismus, Antisemitismus, Politik).

Natur: Beschreibungen von Natur, Naturphänomenen, Flora und Fauna, nicht aber das Senden von (Trocken-)Blumen und Blättern im Brief (siehe Geschenke), nicht die Beschreibung von geografischen Landschaften, wie Meer, Berge (siehe Landschaften), und nicht die Beschreibung von Jahreszeiten oder Wetterphänomenen (siehe Zeit, Wetter).

Öffentliche Räume: Umfasst alle öffentlichen Gebäude und Plätze, wie Gaststätte, Waschhaus, Friedhof, Hafen, Marktplatz, Rathaus, nicht aber der Besuch einer Gaststätte (siehe Freizeit), Reichstag (siehe Politik), Markt (siehe Hausarbeit), Schreibstube (siehe Arbeit) und Truppenunterkunft (siehe Private Räume).

Oper: Umfasst auch Operette, Opernsänger und Opernsängerinnen, den Besuch einer Oper.

Paarbeziehung: Umfasst die engere Beziehung der Schreibenden und anderer Liebes- und Ehepaare, nicht aber die Beziehung zu weiteren Personen (siehe Familie, Freunde, Bekannte).

Partei: Umfasst die spezifischen Institutionen und Organisationen der Nationalsozialistischen Deutschen Arbeiterpartei (NSDAP) und ihrer Ideen, sowie damit verbundener Organisationen, wie Schutzstaffel (SS), Nationalsozialistischer Volkswohlfahrt (NSV), NS-Frauenschaft, Kraft durch Freude (KdF), Hitlerjugend (HJ), Bund Deutscher Mädel (BDM), Organisation Todt (OT), Winterhilfswerk (WHW) und anderer, sowie gleichgeschalteter oder die Ideologie affirmierender Gruppen und Institutionen, wie Bunkergemeinschaft, Ortsvorsteher, Deutsches Rotes Kreuz, nicht aber konkrete Personen (siehe Machthaber) oder die nationalsozialistische Ideologie (siehe Nationalsozialismus).

Politik: Alle Beschreibungen von politischen (und privaten) Handlungen und Entscheidungen, wie Erlasse, Verwaltungsentscheidungen (sofern sie nicht explizit nationalsozialistisch waren), wie Tagespolitik, Vaterland, Siedlungspolitik, Generalplan Ost, Umsiedlungen, Außenpolitik, Kinderlandverschickung, Denunziation, sowie politischen Ideen, wie Nationalismus, Kommunismus, nicht aber die nationalsozialistische Ideologie, nationalsozialistische Gesetze und die Verfolgung von Andersdenkenden (siehe Nationalsozialismus, Antisemitismus, Rassismus, Kriegs-/Volksgemeinschaft).

Praktiken: Handlungen und Verhaltensweisen im Alltag, sowie ihre Bewertungen, wie Ehrgeiz, Eigensinn, Entschuldigungen, Schikane, Gastfreundschaft, Betrug, Schlafen, Mord, Entschluss, Hektik, Selbstreflexion.

Private Räume: Umfasst alle privaten Räumlichkeiten und Gebäude, wie Haus, Garten, Zimmer, Stube, Truppenunterkunft, Schlafsaal, Toilette.

Rassismus: Umfasst Ideen und Handlungen auf Basis der Unterteilung von Menschen nach der zeitgenössischen Idee von Kriterien und Wertigkeiten von Menschen nach kulturellen und biologischen Eigenschaften, wie die Ausgrenzung von Sinti und Roma, Euthanasie gegen Menschen, die als 'minderwertig' klassifiziert wurden, also auch die zeitgenössischen Begriffe Volkstum, Wesensart, Rasse, Rassenpolitik, Rassenbiologie und Rassenhygiene, Rassenkunde, Ahnenforschung, Hautfarbe, nicht aber die Ausgrenzung von Juden und Jüdinnen aus der Gesellschaft (siehe Antisemitismus) und ihre Ermordung in Konzentrations- und Vernichtungslagern (siehe Gräueltaten) und nicht die Ausbeutung von Arbeitskraft aus den besetzten Gebieten (siehe Wirtschaft).

Reden: Umfasst politische Reden im Radio, in der Zeitung, in Schriften, sowie informative Vorträge und Vortragsreihen, nicht aber Predigten (siehe Kirche).

Rundfunk: Umfasst auch Radioapparat, (gemeinsames) Hören von Radiosendungen, Rundfunknachrichten, Wunschkonzert, nicht aber die Inhalte einer politischen Radioansprache (siehe Reden).

Schreiben: Umfasst alle Gedanken, Handlungen und Gegenstände im Bereich des Briefeschreibens und Tagebuchschreibens, sowie den postalischen Verkehr, die Dauer von Postsendungen, Eingriffe in das Schreiben (wie Selbstzensur und Zensur), Briefschulden, Telegramme, Postkarten, Briefpapier, Schreibgerät, nicht aber die Schreibstube (siehe Arbeit).

Sexualität: Umfasst alle Gedanken, Normen und Handlungen im Bereich von Lust, Sexualität, auch Sittlichkeit, Menstruationskalender, Metaphern, sowie weitergefasste Begriffe von Intimität und Nähe, die zeitgenössisch der Kommunikation über Geschlechtsverkehr dienten, sowie Verstöße gegen Sexualnormen und Bewertungen von Sexualität, wie sexuelle Attraktivität, Austausch von intimen Zärtlichkeiten, außereheliche Beziehungen, Geschlechtsverkehr, Kondome, Prostitution, Bordellbesuche, Vergewaltigungen, nicht aber kriegsbezogene Vergewaltigungen (siehe Gräueltaten), und nicht allgemeinere Erwartungen an Geschlechter (siehe Geschlechterrollen).

Sport/Tanz: Umfasst auch Tanzen, Volkstänze, Wandern, Bergsteigen, nicht aber Spaziergang (siehe Freizeit, Tourismus).

Sprachen: Umfasst die Beschreibung von Sprechen, Fremdsprachen, Dialekte, Soldatensprache, Metaphern, Sprachveränderung, nicht aber der schriftlichen Variante (siehe Schreiben).

Status: Umfasst beruflichen und persönlichen Status, auch Stand, Standesdünkel, Bewertungen von Niveau.

Streitkräfte: Umfasst die kämpfenden Einheiten der deutschen und ihrer verbündeten Streitkräfte sämtlicher Truppenteile und Ränge, inklusive des (weiblichen) Gefolges, aber auch der alliierten Streitkräfte und der Zivilverteidigung, nicht aber einzelne Waffengattungen und Ausstattung (siehe Waffen) und nicht der militärische Habitus und Alltag von Soldaten an der Front oder in der Etappe (siehe Militär, Arbeit).

Theater: Umfasst auch Theaterbesuche, Theaterschauspieler und -schauspielerinnen, Fronttheater, nicht aber Oper und Film.

Tourismus: Umfasst alle Gedanken und Handlungen, die den Besuch eines anderen Ortes oder Landes, auch an der Front und in der Etappe, als Reiseerfahrung fasst, wie das Flanieren an Promenaden, das Besorgen von Souvenirs, Nehmen von Erinnerungsfotos, ziellose Umherwandern auf Spaziergängen in der Umgebung, Besuchen von Sehenswürdigkeiten und Erkunden von örtlichen Besonderheiten durch Ausflüge.

Urlaub: Umfasst Urlaubstage, Urlaubsplanung, Urlaubsanträge, Warten auf Urlaub im Sinne einer längeren Pause von der Arbeit oder dem Dienst, meistens verbunden mit einem Besuch in der Heimat, nicht aber das Erkunden der Umgebung eines fremden Landes (siehe Tourismus).

Vergangenheit: Umfasst die Thematisierung von Erinnerungen, vergangenen Ereignissen und Aussagen und auch Ausführungen zu Geschichte.

Waffen: Umfasst Waffengattungen wie Panzer, Seeflotte, und Mittel der Totalisierung des Krieges und Mobilisierung.

Werte: Werte, Normen, Moralvorstellungen, Bewertungen und Gedanken zu (idealem) Verhalten von Menschen und Gesellschaften.

Wetter: Umfasst alle Beschreibungen von Wetter, Wetterphänomen und Wetterfolgen, nicht aber der Jahreszeiten (siehe Zeit).

Wirtschaft: Umfasst Beschreibungen der ökonomischen Lage und Politik in Deutschland und den besetzen Gebieten, wie auch Hamstern und die Ausbeutung von Arbeitskräften durch Zwangsarbeit, nicht aber die persönlichen Praktiken des Hauswirtschaftens und Einnahmen (siehe Geld).

Wissenschaft: Umfasst alle Beschreibungen von wissenschaftlichen Erkenntnissen und Tätigkeiten, wie Astronomie, Theologie, Ethnologie, Psychologie, Philosophie und andere.

Zeit: Umfasst Begriffe der Zeitlichkeit, der Jahreszeiten, des Kalenders, der Wochentage, nicht aber Erinnerungen (siehe Vergangenheit) oder Zukunftspläne (siehe Zukunft).

Zeitungen: Umfasst Zeitungen, Zeitschriften, Wochenblätter, Illustrierte.

Zukunft: Umfasst politische Zukunftsentwürfe und persönliche Zukunftspläne, Wünsche, Fantasien, nicht aber Träume vom Sieg (siehe Kriegsverlauf).

A.3 Count of Keywords in the Dataset

This table reports the counts of each keyword in each part of the dataset.

Train: Split of the labelled data used for training
Dev: Split of the labelled data used for comparing parameters
Test: Split of the labelled data used to test performance
Existing: Labelled part of the dataset (sum of train, dev, test)
Assigned: Labels that were assigned to the unknown data by the classifier
Complete: Sum of existing and assigned labels

Keyword	Train	Dev	Test	Existing	Assigned	Complete
Allianzen	2	1	3	6	0	6
Alter	27	9	6	42	9	51

Antisemitismus	8	1	4	13	0	13
Arbeit	185	38	49	272	400	672
Aus-/Bildung	114	13	9	136	58	194
Ausweise	2	5	2	9	1	10
Baukunst	4	3	3	10	0	10
Begegnungen	217	14	11	242	144	386
Behörde	20	14	3	37	1	38
Bekannte	16	8	13	37	5	42
Bekleidung	53	5	7	65	15	80
Bevölkerung	4	2	4	10	0	10
Essen/Trinken	104	27	36	167	230	397
Familie	390	54	54	498	563	1061
Feste	215	40	73	328	339	667
Film	48	24	17	89	66	155
Fotografie	108	14	10	132	98	230
Freizeit	35	17	10	62	14	76
Freunde	74	2	6	82	15	97
Führer	14	5	7	26	39	65
Gefühle	586	112	128	826	986	1812
Geld	87	7	22	116	145	261
Geschenke	68	19	28	115	80	195
Geschlechterrollen	135	35	31	201	121	322
Gesundheit	119	60	28	207	232	439
Glaube	309	61	58	428	697	1125
Gräueltaten	17	5	1	23	10	33
Hausarbeit	84	39	50	173	209	382
Hausrat	21	8	16	45	11	56
Heimat	93	15	7	115	247	362
Hygiene	73	15	11	99	71	170
Kameraden	118	34	27	179	260	439
Kinder	121	25	30	176	280	456
Kirche	102	22	25	149	167	316
Kommunikation	88	6	9	103	65	168
Kriegs-/ Volksgemeinschaft	20	9	7	36	51	87
Kriegsfolgen	17	11	8	36	1	37
Kriegsschauplatz	28	17	14	59	3	62
Kriegsverlauf	156	60	46	262	504	766
Kultur	25	5	5	35	9	44
Kulturkontakt	33	6	2	41	18	59
Kunst	13	6	7	26	17	43
Körper	18	7	5	30	2	32
Landschaften	24	4	2	30	20	50
Lebenszyklus	41	8	11	60	29	89
Literatur	87	22	24	133	178	311
Luftkrieg	77	9	9	95	130	225
Machthaber	18	1	13	32	2	34
Militär	101	14	28	143	143	286
Mobilität	260	32	27	319	434	753
Musik	102	32	20	154	188	342
Nationalsozialismus	46	30	24	100	9	109
Natur	71	9	20	100	102	202
Öffentliche Räume	103	17	18	138	112	250
Oper	12	1	2	15	15	30

Paarbeziehung	141	13	27	181	203	384
Partei	34	23	16	73	22	95
Politik	65	14	23	102	66	168
Praktiken	46	8	5	59	24	83
Private Räume	29	18	15	62	18	80
Rassismus	17	8	7	32	5	37
Reden	9	4	8	21	12	33
Rundfunk	71	9	5	85	97	182
Schreiben	300	90	86	476	477	953
Sexualität	100	14	13	127	84	211
Sport/Tanz	73	14	6	93	41	134
Sprache	17	3	1	21	1	22
Status	31	1	2	34	1	35
Streitkräfte	104	15	11	130	239	369
Theater	30	5	6	41	17	58
Tourismus	64	27	24	115	45	160
Urlaub	103	27	23	153	251	404
Vergangenheit	79	19	20	118	59	177
Waffen	31	3	3	37	20	57
Werte	119	27	21	167	257	424
Wetter	237	38	37	312	327	639
Wirtschaft	132	89	61	282	273	555
Wissenschaft	33	5	3	41	15	56
Zeit	146	20	35	201	191	392
Zeitungen	7	7	4	18	14	32
Zukunft	199	31	31	261	415	676

Appendix B

Model Training

B.1 Stop Word List

The basic version of the stop word list was taken from spaCy (Montani et al. 2020).[1]

B.1.1 Stop Word List Extension 1

The first extension of the stop word list includes stop words specific to the Oberfrohna corpus:
[“ach”, “oh”, “o”, “schon”, “ganz”, “muß”, “ja”, “gar”, “heute”, “immer”, “januar”, “februar”, “märz”, “april”, “mai”, “juni”, “juli”, “august”, “september”, “oktober”, “november”, “dezember”, “montag”, “dienstag”, “mittwoch”, “donnerstag”, “freitag”, “sonnabend”, “sonntag”, “sonnabendmorgen”, “daß”, “mal”, “womöglich”, “unsr”, “sooo”, “all”, “unsre”, “wo”, “soo”, “habn”]

B.1.2 Stop Word List Extension 2

The second extension of the stop word list includes nicknames the spouses used to refer to each other or, occasionally, to themselves:
[“lieb”, “geliebte”, “liebe”, “lieben”, “herzlieb”, “herz”, “liebste”, “herzallerliebste”, “geliebter”, “liebster”, “liebste”, “herzallerliebster”, “geliebtes”, “herzliebes”, “geliebtester”, “hilde”, “roland”, “herzelein”, “liebstes”, “mannerli”, “herzliebe”, “herzensweibe”, “weiblein”, “schätzeli”]

B.2 Detailed Results of the Stacked Classifier

The stacked classifier is the model that achieved the best performance. The following table reports its performance on each individual class:

Keyword	Precision	Recall	Micro Average	Occurrence in entire labelled data
Allianzen	0	0	0	6
Alter	0.5	0.167	0.25	42
Antisemitismus	0	0	0	13
Arbeit	0.258	0.327	0.288	272
Aus-/Bildung	0.167	0.111	0.133	136
Ausweise	0	0	0	9
Baukunst	0	0	0	10
Begegnungen	0.062	0.091	0.074	242

[1] https://github.com/explosion/spaCy/blob/master/spacy/lang/de/stop_words.py

Behörde	0	0	0	37
Bekannte	0	0	0	37
Bekleidung	1	0.143	0.25	65
Bevölkerung	0	0	0	10
Essen/Trinken	0.3	0.5	0.375	167
Familie	0.297	0.704	0.418	498
Feste	0.482	0.562	0.519	328
Film	0.455	0.588	0.513	89
Fotografie	0.08	0.2	0.114	132
Freizeit	0.333	0.1	0.154	62
Freunde	0	0	0	82
Führer	0.5	0.286	0.364	26
Gefühle	0.527	0.922	0.67	826
Geld	0.286	0.455	0.351	116
Geschenke	0.186	0.286	0.225	115
Geschlechterrollen	0.375	0.29	0.327	201
Gesundheit	0.419	0.464	0.441	207
Glaube	0.307	0.879	0.455	428
Gräueltaten	0	0	0	23
Hausarbeit	0.395	0.34	0.366	173
Hausrat	0	0	0	45
Heimat	0.089	0.714	0.159	115
Hygiene	0.182	0.364	0.242	99
Kameraden	0.255	0.519	0.341	179
Kinder	0.373	0.633	0.469	176
Kirche	0.379	0.44	0.407	149
Kommunikation	0.125	0.111	0.118	103
Kriegs-/ Volksgemeinschaft	0.182	0.286	0.222	36
Kriegsfolgen	0	0	0	36
Kriegsschauplatz	0	0	0	59
Kriegsverlauf	0.293	0.63	0.4	262
Kultur	0	0	0	35
Kulturkontakt	0	0	0	41
Kunst	0.333	0.143	0.2	26
Körper	0	0	0	30
Landschaften	0	0	0	30
Lebenszyklus	0.286	0.182	0.222	60
Literatur	0.121	0.167	0.14	133
Luftkrieg	0.333	0.111	0.167	95
Machthaber	0	0	0	32
Militär	0.176	0.107	0.133	143
Mobilität	0.231	0.556	0.326	319
Musik	0.162	0.3	0.211	154
Nationalsozialismus	0.667	0.083	0.148	100
Natur	0.158	0.3	0.207	100
Öffentliche Räume	0.125	0.167	0.143	138
Oper	0.5	0.5	0.5	15
Paarbeziehung	0.151	0.296	0.2	181
Partei	0	0	0	73
Politik	0.308	0.174	0.222	102
Praktiken	0	0	0	59
Private Räume	0	0	0	62
Rassismus	0	0	0	32

Reden	0	0	0	21
Rundfunk	0.091	0.4	0.148	85
Schreiben	0.4	0.488	0.44	476
Sexualität	0.097	0.231	0.136	127
Sport/Tanz	0.083	0.167	0.111	93
Sprache	0	0	0	21
Status	0	0	0	34
Streitkräfte	0.083	0.182	0.114	130
Theater	0.25	0.167	0.2	41
Tourismus	0.273	0.125	0.171	115
Urlaub	0.212	0.478	0.293	153
Vergangenheit	0.176	0.15	0.162	118
Waffen	0	0	0	37
Werte	0.18	0.429	0.254	167
Wetter	0.25	0.459	0.324	312
Wirtschaft	0.466	0.557	0.507	282
Wissenschaft	0	0	0	41
Zeit	0.188	0.171	0.179	201
Zeitungen	0	0	0	18
Zukunft	0.167	0.613	0.262	261