University of Zurich[UZH]

Master's thesis
for the degree of
Master of Arts
presented to the Faculty of Arts and Social Sciences
of the University of Zurich

Automatic Cluster Analysis of Texts in Simplified German

Author: Battisti Alessia
Student ID: 16-754-192

Examiner: Prof. Dr. Martin Volk

Supervisor: Dr. Sarah Ebling

Institute of Computational Linguistics

Submission date: 14 June 2019

Abstract

Text simplification is the process of reducing lexical and syntactic complexity of a text, while preserving most of the original information content [Saggion, 2017, 1]. This process aims at making texts accessible for everyone, including persons with low literacy skills, cognitive or learning disabilities, aphasia or dementia, among others. Because of the heterogeneity of the target users, simplified German as an instance of simplified language has been conceptualised at multiple complexity levels [Bredel and Maaß, 2016; Bock, 2014; Kellermann, 2014]. However, to date neither guidelines nor evidence support this claim.

In this master thesis, I present an approach to automatically analyse existing texts in simplified German, with the goal of investigating evidence of multiple complexity levels. This approach was tested with two different corpora in simplified German. The first task in my analysis is to address a key question in text simplification research, namely the identification of complexity structures of given texts. This includes the creation of a feature framework reflecting the linguistic and structural characteristics of texts in simplified German. The second task is to cluster documents by exploring various unsupervised algorithms and combinations of the previously extracted features. In the third task, the output of the cluster analysis is validated to calculate its robustness; finally, the clustering results are linguistically interpreted to identify feature behaviours. The results show that clustering techniques are able to discriminate among texts in simplified German, suggesting that some groups of texts share a high degree of linguistic similarity. This thesis emphasises the necessity of exploring not only linguistic features but also structural and layout characteristics of simplified language in order to meet the requirements of the various target users.

Zusammenfassung

Textvereinfachung ist der Prozess der Reduzierung der lexikalischen und syntaktischen Komplexität eines Textes unter Beibehaltung des ursprünglichen Informationsgehalts [Saggion, 2017, 1]. Ziel dieses Prozesses ist es, Texte für alle zugänglich zu machen, auch für Personen mit geringen Lesekompetenzen. Diese Gruppe umfasst unter anderem Menschen mit kognitiven Behinderungen, Menschen mit Lernbehinderungen, Aphasie oder Demenz. Aufgrund der Heterogenität der Zielgruppe wurde leichtes Deutsch als Instanz leichter Sprache auf mehreren Komplexitätsstufen konzipiert [Bredel and Maaß, 2016; Bock, 2014; Kellermann, 2014]. Allerdings stützen bis anhin weder Richtlinien noch empirische Befunde diese Behauptung.
In dieser Masterarbeit präsentiere ich einen Ansatz zur automatischen Analyse bestehender Texte in leichtem Deutsch, mit dem Ziel, empirische Evidenz für das Vorhandensein mehrerer Komplexitätsstufen zu finden. Dieser Ansatz wurde mit zwei verschiedenen Korpora in leichtem Deutsch getestet. Die erste Aufgabe meiner Analyse ist die Adressierung einer Schlüsselfrage der Textvereinfachungsforschung, nämlich die Identifizierung von Komplexitätsstrukturen gegebener Texte. Dazu gehört die Erstellung eines Merkmalsrahmens, der die sprachlichen und strukturellen Merkmale von Texten in leichtem Deutsch widerspiegelt. Die zweite Aufgabe besteht darin, Dokumente in Gruppen einzuteilen, indem verschiedene Algorithmen des unüberwachten maschinellen Lernens und Kombinationen der zuvor extrahierten Merkmale untersucht werden. In der dritten Aufgabe wird das Ergebnis der Clusteranalyse validiert, um ihre Robustheit zu berechnen. In der letzten Aufgabe werden die Clustering-Ergebnisse sprachlich interpretiert, um das Merkmalsverhalten zwischen den Clustern zu identifizieren. Die Ergebnisse zeigen, dass Clustering-Techniken in der Lage sind, Texte in leichtem Deutsch zu unterscheiden, was darauf hindeutet, dass einige Textgruppen einen hohen Grad an sprachlicher Ähnlichkeit aufweisen. Diese Arbeit betont die Notwendigkeit der Erforschung nicht nur linguistischer, sondern auch struktureller und Layout-Merkmale der leichten Sprache, um die Anforderungen der verschiedenen Zielgruppen zu erfüllen.

Acknowledgement

In this two-year-long journey, motivation and curiosity have played a great role and have brought me to this point, in which I am writing the acknowledgement of my thesis. Yet this accomplishment would not have been possible without the help, encouragement and company from several people, whom I would like to dedicate some words.

Firstly, I express my deepest gratitude to my supervisor Dr. Sarah Ebling for introducing this fascinating topic to me. She has been offering her feedback, valuable expertise and insightful criticism, whenever I needed help and advice in regard to this thesis. I am also forever grateful to Prof. Dr. Martin Volk, who unknowingly awakened my interest for simplified language during one of his lecture I attended in the first year.

I would like to thank each institution, translation agency and organisation, especially *capito*, for sharing their knowledge and data in simplified German. In addition, I would like to thank Julia Suter for kindly sharing her technical expertise in analysing simplified German. I am forever thankful to my cousin Marika for sharing her knowledge in Psychology and for the discussions we had on reading comprehension and image perception of children with Autism Spectrum Disorder. A big thank you goes also to all the people who have patiently proofread my English. All remaining errors are my own.

I truly thank Yu for his thoughtful and clever comments on my work and for encouraging me to believe in my potentiality. My special thank goes to my friend and fellow student Isabel, who encouraged me to take on this master project following my passion for e-accessibility and motivated me throughout our study at the University of Zurich.
Last but not least, I am grateful to my parents and my sister in Italy and, of course, to Nadine, who has patiently supported me in good and in bad times.

The first is the paradise in which humans were fully integrated into nature. The second is the artificial paradise, developed by human intelligence to globalizing proportions through science and technology. [...] The Third Paradise is the great myth that leads everyone to take *personal responsibility* in the *global* vision.

— Michelangelo Pistoletto, *The Third Paradise*, 2003

Italian painter, action and object artist, and art theorist

Contents

Abstract i

Acknowledgement iii

Contents v

List of Figures vii

List of Tables ix

List of Acronyms x

1 Introduction 1

1.1 Motivation 1
1.2 Research Questions 3
1.3 Contribution of the Thesis 4
1.4 Thesis Structure 4

2 Research Background 6

2.1 Automatic Analysis of Simplified Texts 6
2.2 Automatic Text Simplification 7
2.3 Automatic Readability Assessment 9
2.3.1 Features for Readability Assessment 11
2.3.2 Further Relevant Features 14
2.4 Summary 18

3 Clustering Simplified German Texts 19

3.1 Data 20
3.1.1 Corpus of Simplified German 20
3.1.2 "TopEasy News" Text Collection 26
3.2 Features 27
3.2.1 Feature Design 27
3.2.2 Feature Transformation 29
3.2.3 Feature Engineering 29

3.3 Evaluation Metrics . . . 30
3.4 Experiment 1: Hierarchical Clustering . . . 31
3.4.1 Method . . . 32
3.4.2 Results . . . 35
3.4.2.1 Without Feature Agglomeration . . . 35
3.4.2.2 With Feature Agglomeration . . . 42
3.5 Experiment 2: K-means . . . 45
3.5.1 Method . . . 46
3.5.2 Results . . . 47
3.5.2.1 Without Feature Agglomeration . . . 47
3.5.2.2 With Feature Agglomeration . . . 54
3.6 Experiment 3: Clustering TopEasy News Text Collection . . . 57
3.6.1 Method . . . 57
3.6.2 Results . . . 58
3.6.2.1 Without Feature Agglomeration . . . 58
3.6.2.2 With Feature Agglomeration . . . 61
3.7 Summary . . . 62

4 Discussion **64**
4.1 Research Question 1 . . . 64
4.2 Research Question 2 . . . 69
4.3 Research Question 3 . . . 70

5 Conclusion **76**
5.1 Conclusion . . . 76
5.2 Future Work . . . 77

Glossary **79**

References **80**

Curriculum Vitae **93**

A Example of a TCF file **94**

B Topic Modelling Analysis **97**

C Complete Feature Set **99**

D Dendrograms **101**

E Tools and Resources **103**

List of Figures

1 Example from LanguageTool . . . 10
2 Example from Prüftool . . . 11
3 Pipeline of the clustering analysis . . . 19
4 Pipeline of the corpus creation . . . 21
5 Example of a text in standard German and simplified German . . . 25
6 Sample of a truncated tree . . . 32
7 Plot of the elbow method: Subsets 1, 3 and 5 . . . 36
8 Example of a text for each cluster in the analysis of the complete dataset: Subset 2 . . . 37
9 Example of texts from the analysis of PDFs. . . . 40
10 Silhouette profile of two clustering models . . . 41
11 Complete dendrogram of webpages . . . 41
12 Complete dendrogram of PDFs . . . 42
13 Truncated tree after feature agglomeration for PDFs . . . 45
14 Plots of the evaluation metrics using k-means . . . 47
15 Plot of the silhouette coefficient and scatter plot using k-means . . . 48
16 Scatter plot using k-means: Subset 2 . . . 49
17 Example of a text for each cluster for the PDFs: Subset 2 . . . 50
18 Plots of the evaluation metrics across feature subsets using k-means . 51
19 Plot of the silhouette coefficient and scatter plot of PDFs. . . . 53
20 Plots of the elbow method using k-means . . . 55
21 Scatter plots of using k-means after agglomerating Subset 1 . . . 56
22 Plots of elbow method on “TopEasy News” text collection . . . 59
23 Plots of the silhouette scores for 2, 3 and 4 clusters of “TopEasy News” text collection . . . 60
24 Scatter plot of “TopEasy News” text collection using k-means . . . 60
25 Box plots of Subset 2 . . . 65
26 Box plots of Subset 4 . . . 66
27 Box plots of Subset 5 . . . 67
28 Box plots of Subset 3: webpages vs. PDFs . . . 68
29 Correlation matrix of Subset 4 . . . 72
30 Example of a text at level B1 alongside with its A2 counterpart . . . 73

31 Plot of agglomerated features of Subset 2 73
32 Plot of agglomerated features of Subset 1 75
33 Visualisation of dendrograms for the complete dataset 101
34 Visualisation of dendrograms for “TopEasy News” text collection . . 102

List of Tables

1 Morphological, morpho-syntactic and syntactic features 13
2 Lexical and semantic features 15
3 Relevant features 17
4 Corpus profile: PDF vs. HTML 23
5 Complete corpus profile 24
6 Complete profile of "TopEasy News" text collection 27
7 Feature engineering 30
8 Comparison of the CCC results 33
9 Comparison of evaluation metric scores in experiment 1 35
10 Comparison of evaluation metric scores for webpages and PDFs in experiment 1 38
11 Comparison of the evaluation metrics after feature agglomeration in experiment 1 43
12 Comparison of the evaluation metrics after feature agglomeration for webpages and PDFs in experiment 1 44
13 Comparison of the evaluation metrics on webpages and PDFs in experiment 2 53
14 Comparison of the evaluation metrics on webpages and PDFs after feature agglomeration in experiment 2 56
15 Comparison of the evaluation metrics among the five feature subsets in experiment 3 58
16 Comparison of the evaluation metrics among the five feature subsets in experiment 3 61
17 Comparison of the evaluation metrics for hierarchical clustering and k-means after feature agglomeration in experiment 3 62
B1 Topics NNMF 98
B2 Topics LDA 98
C1 Complete feature set. 99

List of Acronyms

ASD	Autism Spectrum Disorder
ARA	Automatic Readability Assessment
ATS	Automatic Text Simplification
BITV	*Barrierefreie-Informationstechnik-Verordnung*
CCC	Cophenetic Correlation Coefficient
CEFR	Common European Framework of Reference for Languages
CMDI	Component MetaData Infrastructure
HMTL	HyperText Markup Language
LDA	Latent Dirichlet Allocation
LIX	*Läsbarhetsindex*
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Processing Toolkit
NMT	Neural Machine Translation
NNMF	Non-Negative Matrix Factorisation
OCR	Optical Character Recognition
OG	Original German
OLAC	Open Language Archives Community
PBMT	Phrase Based Machine Translation
PCA	Principal Component Analysis
POS	Part Of Speech
SG	Simplified German
SMT	Statistical Machine Translation
SSE	Sum of Squared Errors
TCF	Text Corpus Format
TF-IDF	Term Frequency-Inverse Document Frequency
XML	eXtensible Markup Language

1 Introduction

1.1 Motivation

The research field of automatic text simplification (ATS) studies methods and techniques to automatically simplify textual content. Simplified texts have been suggested as a potential easier-to-handle input for text processors, such as parsers or machine translation tools [Bernth, 1998]. An alternative purpose of simplified language is providing accessible, readable and understandable information to persons with reduced reading abilities [Siddharthan, 2014]. This group includes persons with cognitive disabilities, learning disabilities, dementia, aphasia, prelingual deafness and functionally illiterate persons. Non-native speakers, such as language learners and immigrants, and children can be also considered as target users of simplified language, but their difficulty in accessing complex texts is expected to improve. The Swiss Federal Statistical Office estimates that in Switzerland more than 1.4 million persons face a disability. Among them, 800,000 persons, namely 16% of the population between the ages of 16 and 65, are dealing with reading difficulties [Notter et al., 2006][1]. Simplified language is characterised by reduced lexical and syntactic complexity and includes images, structured layout and explanations for difficult words. Throughout this thesis, I will use the term *simplified language* because it refers to a more general concept of complexity reduction. Other terms, such as *plain language*, concern a specific level of simplification. The adjectives in the expressions *easy-to-read language* or *simple language* imply that this type of language is straightforward and comprehensible, which is clearly not the case for the target readers.

Various countries have acknowledged simplified language as a means of inclusion that enables the target population mentioned above to inform themselves of their legal rights and participate in society. While Nordic countries have pioneered the promotion of simplified language for decades, little research has been conducted in this field for the German language. In Germany, official communications on government-level websites have been declared in simplified German only since 2014

[1] An update of this statistics is going to be published in 2022 in the context of OECD-Survey: `http://www.oecd.org/skills/piaac/` (last accessed: 11 April 2019)

according to a set of regulations for accessible information technology (*Barrierefreie-Informationstechnik-Verordnung*, BITV 2.0) [Bundesministerium für Arbeit und Soziales, 2011]. Published in 2006, Austria and Switzerland approved the United Nations Convention on the Rights of Persons with Disabilities (CRPD) in 2008 and 2014, respectively. The CRPD explicitly includes "plain-language" in the Communication definition [United Nations 2006, Article 2] and wishes to "promote access for persons with disabilities to new information and communications technologies and systems, including the Internet" [United Nations 2006, Article 9, Paragraph 2.g].
Since then, a large number of texts in simplified German has become available on the web. For German, several guidelines provide recommendations for writing in simplified language, defining which structures need to be avoided, which need to be paraphrased and which are understandable. Four guidelines are worth mentioning: the rules published by Inclusion Europe [2009], the BITV 2.0 [2011], the rules of the Netzwerk Leichte Sprache [2013] and the rules written by Maaß [2015]. Among the total 120 rules in these guidelines, only 17 overlap [Bredel and Maaß, 2016], demonstrating that simplified language is not a standardised concept. Furthermore, these rules do not consider the heterogeneity of the target groups.

Only few authors have conceptualised simplified language as a construct with multiple complexity levels. What is more, no guidelines have been proposed to define and implement these levels. Bock [2014] classifies simplified German into three levels: *leichte Sprache*, *einfache Sprache* and *bürgernahe Sprache*; the latter corresponds to a specific variety of simplified language for technical languages and terminology [Bredel and Maaß, 2016]. Kellermann [2014] distinguishes between *leichte Sprache* for persons with cognitive disabilities and *einfache Sprache* for foreign language learners, elderly people, etc. Bredel and Maaß [2016] define *einfache Sprache* as an enriched form of *leichte Sprache* and hence suggest adding more information to this variety. Considering the Common European Framework of Reference for Languages (CEFR) [Council of Europe, 2001], *leichte Sprache* is typically associated with level A1, while *einfache Sprache* corresponds to levels A2 to B1. Similarly, the social franchise network *capito*[2] recognises three levels of simplified German corresponding to the CEFR levels A1, A2 and B1.

Considering ATS in general, tools for simplifying texts have not convincingly proven to positively and significantly affect text understanding. Saggion [2017] claims that evidence-based research is needed to provide tools that support the developers in creating suitable ATS systems and providing better solutions for real users. Focusing

[2] *capito* provides simplification services and tools, training courses and workshops for primary and secondary recipients of simplified language. `https://www.capito.eu/en/about-us/` (last accessed: 11 April 2019)

on German, Bock [2014] observes a lack of research in simplified German, making it impossible to distinguish between multiple complexity levels.
Within the above framework, this thesis aims to automatically analyse existing simplified texts to empirically search for evidence of multiple complexity levels in simplified German and describe the linguistic characteristics of each level. To the best of my knowledge, this work is the first endeavour to empirically analyse and inductively derive complexity levels for simplified German. Developing a full-fledged framework of complexity levels for the simplified German texts analysed is beyond the scope of this work.

1.2 Research Questions

Departing from the above observations, the key research questions I address in my thesis are the following:

1. Is there empirical evidence of multiple complexity levels in existing simplified German texts?

2. Can an unsupervised machine learning approach, such as cluster analysis, provide such empirical evidence?

3. Are linguistic features suitable for this type of analysis?

Research question 1 is based on the observations presented in Section 1.1, namely that German is simplified on different language levels aiming at different target users. Related to this issue is the exploration of the linguistic, structural and typographic characteristics of the language level(s) identified. Throughout the experiments and in the discussion, I will linguistically examine the results of the cluster analysis to identify some of the most significant and informative features characterising each cluster.

Research question 2 focuses on unsupervised methods. Among various unsupervised approaches, I investigate clustering methods. Due to its exploratory nature, cluster analysis is used to discover natural patterns or grouping in a dataset so that items in the same cluster are more similar to each other than to items from different clusters. Cluster analysis may also result in a single cluster comprising all applicable texts if the characteristics of the documents do not warrant any subdivision of the data. In contrast to supervised learning (e.g., text classification), unsupervised approaches do not consider any ground truth class labels that allow for validating the results.

Research question 3 addresses the main problem of feature engineering, where the

most appropriate feature set for the subsequent analysis is selected. The goal of feature engineering or selection is to come up with the smallest set of features that best capture the characteristics of the problem addressed. In this thesis, I concentrate on linguistic, structural and typographic features looking for a balance between expressiveness and compactness of the overall feature set. Therefore, I combine the features in different ways to find the most representative combinations.

1.3 Contribution of the Thesis

The contributions of this thesis are the following:

1. The identification of multiple levels of complexity and associated characteristics. These characteristics can feed into further research into and development of ATS systems.

2. This thesis tests and discusses the usefulness of structural and typographic features, which – to the best of my knowledge – have never been investigated in clustering or classification analyses related to readability assessment.

3. On a theoretical level, this thesis contributes to a standardisation of simplified German and offers suggestions to improve current guidelines for simplified German. Simplified-language translators may benefit from consulting empirically validated linguistic rules.

4. This thesis serves as the basis for investigations into simplified-language varieties of other languages in that the analysis carried out here can be adapted to other languages, e.g., Italian and French.

1.4 Thesis Structure

The thesis is organised into five chapters. In Chapter 2, I provide an overview of the various tools and applications for analysing and processing simplified documents. I focus primarily on the automatic identification of linguistic structures that might make texts more difficult to understand and on the relevance of selecting appropriate linguistic features, taking complexity into account. In Chapter 3, I describe the core elements of my approach. Firstly, I introduce two new collections of texts in simplified German (Section 3.1). Then, I focus on the process of designing, extracting and transforming the features from the texts, so as to use them as input for the automatic clustering analysis (Section 3.2). In Section 3.3, I introduce the evaluation

metrics I used to evaluate the experimental results. In Sections 3.4, 3.5 and 3.6, I describe my experiments aimed at finding empirical grounding for language levels in simplified German. In Chapter 4, I summarise and discuss the outcomes by exploring the clusters in relation to the data. Chapter 5 concludes this thesis and provides a summary and suggestions for future work in this area.

2 Research Background

In this chapter, I survey research on automatic processing of simplified texts. Firstly, I describe the topic of automatic text simplification (ATS), which nowadays is mostly conceptualised as an instance of machine translation. I then explain different approaches to automatically identify linguistic and structural characteristics as textual indicators of the simple/complex dichotomy. Finally, I describe the classification of simplified texts based on complexity levels and situate the current thesis in the research landscape.

2.1 Automatic Analysis of Simplified Texts

Different techniques for automatic analysis of simplified texts have been explored and developed, focusing either on text simplification (grammatical or lexical) [Specia, 2010], readability assessment [Aluísio and Gasperin, 2010] or both [Vajjala, 2015]. Automatic text simplification aims at making texts easier to read and understand for target readers (cf. Section 2.2). Identifying the complexity of a sentence or text can help to assess if the output of ATS is adequate to the reading ability of a target reader. In this context, the investigation of language complexity is of great importance.
Various studies in linguistic and readability assessment have focused on the question which linguistic features make one language more simple or difficult than others (cf. Section 2.3). Yet readability assessment approaches not only imply language varieties characterised by reduced lexical and grammatical devices but also tend to categorise these varieties into different predetermined complexity levels (e.g., cf. CEFR in Section 1.1). The task of defining what makes a text simple or complex is not straightforward [DeKeyser, 2005], especially when working with text simplification for readers with cognitive impairments.

The work in the context of this thesis can be considered a preliminary stage of the readability assessment task, insofar as I investigate whether complexity levels exist in current simplified German practice. The complete experimental setup, including

feature selection, is described in Chapter 3.

2.2 Automatic Text Simplification

ATS is used to reduce lexical and syntactic complexity, preserving the original semantics of the text. Lexical and syntactic simplification in the context of ATS are usually treated as separate tasks, despite being naturally interconnected. The first task focuses on the identification of difficult words and tries to either replace these words with simpler expressions (Example 2.1) or include suitable explanations (Example 2.2).

(2.1) Olivia is a *bright* girl

Olivia is a *clever* girl.

(2.2) The girl studies *zoology.*

The girl studies *zoology, the science of studying animal life.*

The second task attempts to recognise complex syntactic patterns and convert the sentence into more understandable equivalents (Example 2.3).

(2.3) The speech held by the president was boring.

The president held a speech. The speech was boring.

Early ATS approaches explored hand-crafted rules applied to the output of a syntactic parser [Candido et al., 2009; Chandrasekar et al., 1996; Siddharthan, 2002; Suter, 2015]. The errors of such systems depended on previous parsing or preprocessing errors [Brouwers et al., 2014; Drndarević et al., 2013; Siddharthan, 2011, 2014].

A new direction in ATS research was given by the availability of a large parallel corpus for English and Simple English [Zhu et al., 2010], the first collection of aligned texts from the English[1] and Simple English Wikipedia[2]. This has allowed ATS to be handled as a special case of machine translation (MT), which relies on parallel data in order to translate the source language into a simplified version of the same language. For this reason, ATS is often referred to as "monolingual translation" [Siddharthan, 2014, 13]. Since the publication of the parallel Wikipedia corpus for English, the interest in ATS and developing parallel corpora has grown. This is confirmed by the number of languages (English excluded) for which studies, corpora and systems exist, among which are: Basque [Aranzabe et al., 2012], Brazilian Portuguese [Aluísio

[1]English Wikipedia: `http://en.wikipedia.org/` (last accessed: 11 April 2019)

[2]Simple English Wikipedia: `http://simple.wikipedia.org/` (last accessed: 11 April 2019)

and Gasperin, 2010], Danish [Klerke and Søgaard, 2012], Dutch [Bulté et al., 2018], French [Brouwers et al., 2014; Seretan, 2012], Japanese [Kajiwara and Komachi, 2016; Inui et al., 2003], German [Klaper et al., 2013; Suter et al., 2016], Italian [Barlacchi and Tonelli, 2013; Brunato et al., 2015, 2016], Spanish [Saggion et al., 2011; Sanja et al., 2013] and Swedish [Rennes and Jönsson, 2015].

ATS via MT followed the evolution of MT paradigms. Specia [2010] and Coster and Kauchak [2011] adopted phrase-based machine translation (PBMT), training systems with the Moses toolkit [Koehn et al., 2007]. Siddharthan [2014] argued that PBMT systems were not optimal for reordering and splitting operations and could only perform a small set of simplification operations due to their lack of linguistic knowledge. Zhu et al. [2010] pursued a tree-based statistical machine translation approach (SMT), and Xu et al. [2016] adapted a syntax-based SMT system. Finally, Nisioi et al. [2017], Wang et al. [2016], Zhang and Lapata [2017] applied neural machine translation (NMT), which simultaneously performed lexical simplification and content reduction, mildly outperforming the accuracy of the best phrase-based and syntax-based MT approaches [Nisioi et al., 2017; Stajner and Nisioi, 2018]. For NLP tasks, even a slight improvement in the model accuracy is sufficient for ATS to be applied. However, for user-targeted ATS systems, a low-to-medium model accuracy corresponds to an output more difficult to understand than the original complex text [Saggion, 2017; Shardlow, 2014]. Furthermore, ATS via NMT is currently limited by the scarcity of suitable parallel simplification data.

ATS on different complexity levels and for different target audiences has not been properly developed yet. This issue is strictly linked with the lack of corpora with multiple complexity levels. Exceptions here are the few corpora that have been developed to exemplify the needs of various users of simplified language. Within the PorSimples project[3], Caseli et al. [2009] built a corpus of texts extracted from a newspaper aligned with their simplified versions across two levels. The difference between the two levels consisted of the number of simplified sentences. One level included only some simplified sentences, while on the other level, all sentences were simplified [Gasperin et al., 2009]. The Newsela corpus presented by Xu et al. [2015] contains simplified texts at different complexity levels for children at various educational stages.[4] The corpus includes news articles simplified by professional editors on four levels along with the original texts. Sentence alignment was later performed by Štajner et al. [2017].

[3] `http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources?layout=edit&id=27` (last accessed: 11 April 2019)

[4] `https://newsela.com/data/` (last accessed: 11 April 2019)

2.3 Automatic Readability Assessment

Recognising complex structures in sentences and texts can help to create suitable material for specific target groups and to assess whether a given text matches the reading and understanding ability of a determined target reader. Over the past century, more than 200 readability formulae have been proposed [DuBay, 2004] and criticised [Collins-Thompson and Callan, 2004; Feng et al., 2009]. Readability formulae consider only morphological word complexity (e.g., word length in syllables), sentence complexity features (e.g., sentence length in words) and sometimes require word passages of a fixed size to give a reliable estimate (e.g., 100 words). Hence several more sophisticated readability models have been designed, such as the language model in Cha et al. [2017]. The majority of studies concerning feature engineering and application of readability formulae and models have been conducted for English. Along with this research trend, commercial and non-commercial applications have been developed to automatically identify text properties accounting for text readability. These tools are used to evaluate professionally edited texts (e.g., for English: TextEvaluator[5]) or student essays (e.g., for English: e-Rater[6]), disregarding the assessment of texts written in simplified language for different target groups.

Considering the target users of simplified language, it is necessary to distinguish between readability and understandability [Rello et al., 2013]. Readability is defined as the ease with which a written text can be understood by a reader. Understandability refers to the accessibility level at which a text can be comprehended. Since readability strongly affects text comprehension, both terms have been used interchangeably [Inui et al., 2003]. However, the readability formulae and models presented above are not suitable for assessing texts in simplified language.

LanguageTool [Naber, 2003] is a free and open-source spell and grammar checker for multiple languages that offers a component for testing understandability of texts in simplified German.[7] Given a list of rules, the system detects and highlights which patterns in a given text do not conform to these rules. LanguageTool for simplified German is based on the guidelines published by Netzwerk Leichte Sprache [2013] (Section 1.1) and is available on the LanguageTool and Hurraki websites, a Wiki website for plain-language texts[8]. Figure 1 shows how the system marks seven errors in a simplified text published by Hurraki. Due to the presence of the *Mediopunkt* or *centred dot*, a typographic device proposed by Maaß [2015] for visually segmenting

[5] `https://textevaluator.ets.org/TextEvaluator/` (last accessed: 11 April 2019)

[6] `http://www.ets.org/erater/about` (last accessed: 11 April 2019)

[7] `https://www.languagetool.org/de/leichte-sprache/` (last accessed: 11 April 2019)

[8] `https://hurraki.de/wiki/Hauptseite` (last accessed: 11 April 2019)

German compound words, the system flags the compound *Amtssprache* (en: *official language*) as incorrect. It is important to note that the *Mediopunkt* is neither part of the standard German language nor a sign easily accessible on standard computer and smartphone keyboards. In the first sentence, this error leads the system to also flag the verb *benutzten* (simple past of the verb *benutzen*, en: *to use*), but it does not explain the cause of this detection. In the second sentence, the system suggests to split the sentence at the conjunction *und* (en: *and*) ignoring that this conjunction connects two noun phrases and not two complete clauses. Surprisingly, the system does not identify the named entity *Vereinte Nationen* (en: *United Nations*). Consequently, it reports that the adjective (*Vereinte*) does not match the case required by the preposition *bei* and flags the noun *Nationen* (en: *nations*) as a foreign word.[9]

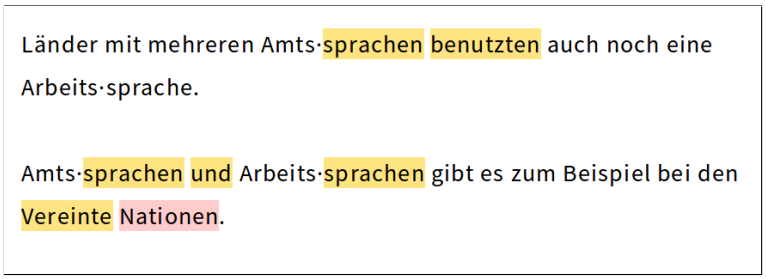


Figure 1: Example of analysis using LanguageTool and a text section taken from the Hurraki entry *Amtssprache*[10].

Designed with the same objective, *capito* (cf. Section 1.1) has been developing a text checker called *Prüftool*. The prototype will be freely available on the web[11] and aims at testing the quality of a text with regard to a precise target group and *capito* level (recall that *capito* departs from the concept of three complexity levels for simplified German: A1, A2 and B1). Before running the test, the user has to answer a number of questions to define the target reader profile, desired language level, purpose of the text and the level of comprehension required by the target reader. As Figure 2 shows, the tool detects and highlights the grammatical and lexical patterns that do not fit the *capito* parameters (top left), as for example that the word *Adoption* (en: *adoption*) is not included in the basic German lexicon. On the right side of the interface, the tool lists the rules that were violated and suggests ideas for improvement. The main goal of this application is to verify the quality of documents written by *capito* translators. However, it can also be applied to assess

[9] In standard German, the adjective *Vereinte* must be declined. The correct prepositional phrase is *bei den Vereinten Nationen*.

[10] `https://hurraki.de/wiki/Amtssprache` (last accessed: 11 April 2019)

[11] The system is currently available only for research.

the quality of texts translated by ATS systems.

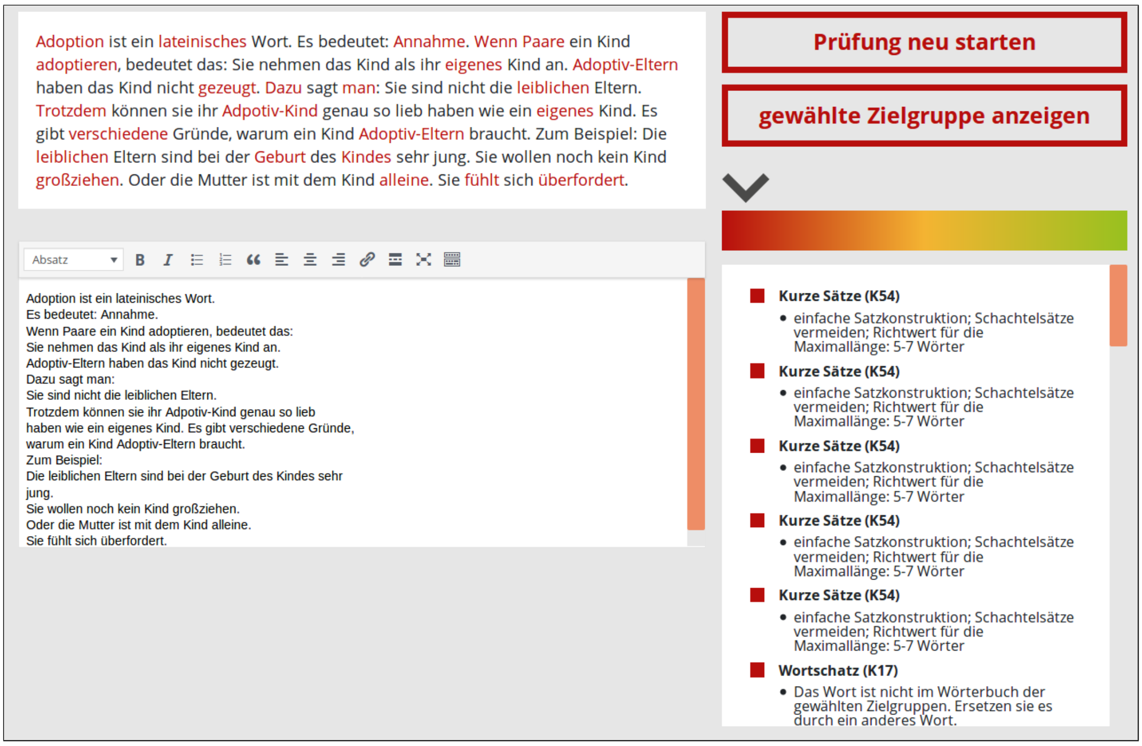


Figure 2: Example of analysis using *capito Prüftool* and a sample text taken from the Lebenshilfe dictionary entry *Adoption*[12].

2.3.1 Features for Readability Assessment

The automatic identification of simple vs. complex structures has also been treated as a classification issue. Classification is a supervised learning approach that sorts and assigns data into predefined classes or categories. The first classification approaches are based on binary representation of the data and only distinguish simple from difficult texts [Hancke et al., 2012; Kauchak et al., 2014]. Recent methods attempt to classify documents at different complexity levels, mostly represented by the CEFR categorisation [Hancke, 2013; Pilan and Volodina, 2018; Reynolds, 2016]. A large number of linguistic features have been proposed to identify complexity in texts. Their selection can help with understanding the main factors making up linguistic complexity and determining whether specific target readers can understand a certain input text at different language levels.

[12] `https://www.lebenshilfe.de/woerterbuch/woerterbuch-detail/?tx_lfdictionary_detail%5Bdictionaryentry%5D=9&tx_lfdictionary_detail%5Bsearchterm%5D=&tx_lfdictionary_detail%5Bletter%5D=A&tx_lfdictionary_detail%5Baction%5D=show&tx_lfdictionary_detail%5Bcontroller%5D=Dictionaryentry&cHash=937b7ea6544cd045998169837e49a532` (last accessed: 11 April 2019)

Surface Features In a language learning (L2) context, count-based measures and syntactic features have been verified to influence L2 complexity [Curto et al., 2015; Reynolds, 2016]. Among count-based features, the index LIX (*Läsbarhetsindex*) represents a general rule of thumb measure that combines the sum of the average number of words per sentence in the text and the percentage of tokens longer than six characters [Björnsson, 1968; Pilan and Volodina, 2018]. Pilan and Volodina [2018] computed the count of punctuation marks, as, in large quantities, their presence could be indicative of a complex syntactic structure. However, the authors did not distinguish between different types of punctuation. For example, unlike commas, the frequent usage of full stops is to be expected in a simplified text or at a low CEFR level and, consequently, indicative of textual simplicity. Further surface features may focus on the character level, as Lau [2006] explored when classifying Chinese texts.

Syntactic Features Graesser et al. [2004] assert that, when reading and processing sentences, syntactic aspects are connected to the reader's working memory load, which can be considerably increased by embedded constituents or syntactic ambiguity. Various studies [Dell'Orletta et al., 2011; Heimann Mühlenbock, 2013; Pilan and Volodina, 2018; Schwarm and Ostendorf, 2005] enumerate features based on dependency parsing, such as the average length of dependency arcs, their direction and the average length and number of noun phrases (NPs), verb phrases (VPs) and prepositional phrases (PPs) per sentence [Feng, 2010; Petersen and Ostendorf, 2009]. The presence of subjunctions and conjunctions can indicate a more complex syntactic structure [Pilan and Volodina, 2018; Suter et al., 2016]. All types of subordinate clauses, relative clauses, pre-modifiers, post-modifiers and prepositional complements are usually considered as indicative of a high degree of complexity [Heimann Mühlenbock, 2013; Schwarm and Ostendorf, 2005]. Analysing Swedish, Pilan and Volodina [2018] and Heimann Mühlenbock [2013] included particles among syntactic features, as they could change the meaning of verbs, just like English phrasal verbs. The readability assessment models implemented for Swedish by Falkenjack and Jonsson [2014] disregarded parsing-based features, yet their models obtained similar results to other readability models.

Morphological Features Morphological features are among the most informative and influential characteristics for morphologically rich languages, such as German [Hancke and Meurers, 2013], Estonian [Vajjala and Lõo, 2014] and Russian [Reynolds, 2016]. In the context of text simplification, Dell'Orletta et al. [2011] take advantage of the rich verbal morphology of Italian, including verbal mood variables in their feature set. François and Fairon [2012] considered not only the verbal mood but also verb tense-based features to create a French readability classification model.

Hancke et al. [2012] developed features based on compounding, which is a productive morphological process of word formation in German [Fleischer, 1975, 85]. They considered the ratio of compound nouns to all nouns and the average number of words in a compound. For Swedish, Pilan and Volodina [2018] included the ratio of a morphological category to the ratio of lexical tokens, such as nouns, verbs, adjectives and adverbs. Table 1 summarises the morphological, morpho-syntactic and syntactic features.

Morphological, Morpho-syntactic and Syntactic Features
Compounds
Verbal Mood
Tense (Present vs. Past or Future)
Coordinating conjunctions
Subjunctions
Pre- and post-modifiers
Particles
Frequency of relative clauses
Participial clauses
Causal clauses
Conditional clauses
Concessive clauses
Final clauses
Interrogative clauses
Modal clauses
Temporal clauses
Parse tree depth
Avg. length of NPs, VPs, PPs

Table 1: Summary table of morphological, morpho-syntactic and syntactic features.

Lexical Features In the language learning context, the most predictive features for simple vs. complex German include not only morphological features but also lexical features. The frequency of words influences lexical complexity, since repeated exposure facilitates their processing [Graesser et al., 2004]. Besides frequency, lexical richness can contribute to distinguishing between difficult and simplified texts. Type-token ratio (TTR), namely the ratio of the number of unique words (types) to the number of occurrences of these words in a text (tokens), is an example of a lexical richness indicator. Vajjala and Meurers [2012] suggested the usage of a bilogarithmic and a square-root TTR to decrease the effect of text and sentence length. In informative and investigative texts, information is expressed with nouns instead

of verbs, making the textual processing more difficult. Nominal ratio [Hultman and Westman, 1977] represents the ratio of nominal categories (nouns, prepositions, participles) to the ratio of verbal categories (verbs, adverbs, pronouns). A simplified version of this formula computes the number of nouns divided by the number of verbs in the same text, still giving a rough idea about the information in a specific text [Heimann Mühlenbock, 2013]. A high noun/pronoun ratio is a further indicator of textual complexity, since pronouns are relevant for cohesion and make texts more difficult to understand [Graesser et al., 2011].

Semantic Features Semantic features include the count of available word senses per lemma based on a given lexicon [Hancke et al., 2012; Pilan and Volodina, 2018]. Semantic ambiguity is demanding for readers, because they need to disambiguate the meaning of individual words while processing the meaning of the whole sentence [Graesser et al., 2004]. This process often results in slower response times and longer processing and fixation times [Kauchak et al., 2014; Rayner and Duffy, 1986].

Cognitively-motivated Features Yaneva et al. [2016] investigated text and web accessibility for persons with Autism Spectrum Disorder (ASD). The authors extracted cognitively-motivated features, including word frequency, age of acquisition of words, words imagability, concreteness and familiarity of words among more classical features, such as lexico-semantic information, superficial syntactic information, cohesion and readability indices. They also included first- and second-person pronominal references as features, since they hypothesised that the number of personal words in a text improves comprehension [Freyhoff et al., 1998]. Entity density, namely the count of entities, such as persons, locations and organisations, and lexical chains (synonymy or hyponymy relations between nouns) are further relevant features to develop for analysing texts targeting persons with mild intellectual disability [Feng et al., 2009; Jansche et al., 2010; Yaneva et al., 2016]. Table 2 shows the introduced lexical and semantic feature set, also containing the cognitively-motivated features of Yaneva et al. [2016].

2.3.2 Further Relevant Features

In this section, I introduce features that have been considered in theoretical studies of simplified German but never implemented in a classification or practical analysis. Focusing on numbers, numerical information can pose comprehension problems for different types of readers [Saggion, 2017]. Roman numerals, numbers written in words, percentages and year dates are attributes of complex texts. Digits and approximate numbers are easier to read than words and exact numbers, respectively.

Lexical and Semantic Features
Word list items (frequent word, complex word, internationalism)
Named entities
Nominal ratio
Noun/pronoun ratio
Bilog and root Type/Token ratio
Lexical density
Lemma variation
Number of proposition per sentence
Entity density
Lexical chains
Word imagability
Word concreteness
Word familiarity
Ambiguity
Word senses

Table 2: Summary table of lexical and semantic features.

In some situations, approximate numbers prevent the reader from having to process and understand unnecessary information [Bautista and Saggion, 2014]. Example 2.4 is easier to recognise and read than Example 2.5.

(2.4) almost 2 million words

(2.5) 1,999,983 words

Abbreviations, initial letters and special characters, such as the paragraph symbol §, increase the target readers' working memory load, because explicit knowledge about writing conventions is needed to unlock them [Maaß, 2015]. In the category of special characters, I include the above-mentioned *Mediopunkt* [Maaß, 2015] and hyphens, which are used to split compounds in order to increase readability and disambiguate long words (Examples 2.6[13] and 2.7). The centred dot might lead a system to wrong detection of patterns as I have explained in relation to Figure 1.

(2.6) OG: Handwerker

SG: Hand·werker

[13]In this and the subsequent examples: OG for *original, standard German* and SG for *simplified German*.

(2.7) OG: Arbeiterwohlfahrtversicherungsunternehmen

SG: Arbeiterwohlfahrt-Versicherungsunternehmen

Bredel and Maaß [2016] underline the importance of a structured layout, which clearly renders headings and paragraphs. They also remark on the concept of "one-sentence-per-line", that is, every line in a text represents only one sentence, which contains only one piece of information. Additionally, visible indentation and increased line spacing improve readability, while multiple columns and a small font size tend to make the reading process more difficult. In relation to the font type, a sans-serif font may be easier to read. More recent studies [Sieghart, submitted] challenge this view. Sieghart [submitted] proves that serif fonts allow for a faster decryption of words while reading.

In a simplified text, typographical contrast, such as boldface and italics, serves as a discourse marker signalling words and phrases that require particular attention [Arfé et al., 2018]. These formatting devices on words usually convey different purposes: Boldface marks relevant expressions in the text, while underlining and framing indicate external functions, such as the presence of a hyperlink. McNamara et al. [2012] affirm that signalling as a cognitive text simplification technique helps readers activate background knowledge and aids in text understanding. However, italics change the form of characters and therefore make the reading comprehension process more difficult.

In the psychology of perception, it is known that information is well-processed and absorbed if multi-coded (different types of signals/signs) and multi-modal (different perception channels, i.e., visual and auditive) [Grandin, 1995; Schnotz, 2014]. Images should also activate previous knowledge and exemplify the objects in the text. However, if a target reader's previous knowledge is marginal, images can cause cognitive overload [Sweller et al., 1998; Sweller, 2005].

In semantics, the phenomenon of negation is cognitively demanding since the reader has to imagine and understand the opposite of what is actually written. Similarly, impersonal constructions do not explicitly refer to specific entities in the real word.

On the syntactic level, the standard word order should be followed, namely Subject-Verb-Object (SVO) for German. In a similar way, a subordinate clause must follow the main clause, unless a different sentence order is more understandable, such as to reflect the temporal succession of events.

Morphological and morpho-syntactic features in simplified German encompass the genitive case, genitive attributes, prepositions with genitives, personal pronouns, adverbs, passive constructions and indirect speech. For instance, a genitive attribute

may be changed into a prepositional phrase using the construction *von*+dative (case), which bears the same meaning of possession (Example 2.8).[14]

(2.8) OG: das Buch der Lehrerin
SG: das Buch von der Lehrerin

To summarise, Table 3 shows the miscellaneous features introduced in this section.

Miscellaneous Features
Images
Font type
Font size
Formatting devices
Number of columns
Indentation
Line spacing
Text structure (headings, paragraphs, lines)
Punctuation marks (. ? vs. , ;)
Special characters (*Mediopunkt*)
Numerals (Arabic and digits vs. Roman and numbers in words)
Abbreviations and initial letters
Negation
Impersonal constructions
Question words
Adverbs (Local, temporal, prepositional vs. others)
Pronouns (Demonstrative; 1- and 2-person pronouns vs. others)
Prepositions (Prepositions with genitives, *Wechselpräpositionen*)
Standard word order
Sentence order
Number of main clauses
Attributive relation (*von*+dative or genitive modifiers)
Verbal voice (Active vs. Passive)
Indirect speech

Table 3: Summary table of relevant features for simplified German.

[14]Note that this is a case where simplified German conflicts to some degree with the grammar of standard German, which encourages the former expression.

2.4 Summary

In this chapter, I have introduced common practices for the automatic analysis of simplified texts. Firstly, I have reported on the progress of automatic text simplification. Then, I have described a set of characteristics specific for simplified German and a wide range of surface and deeper linguistic features used in NLP approaches to discover difficult or simple patterns in texts.

The table in Appendix C illustrates the complete list of features introduced in this chapter. The features I implemented for the experiments in Chapter 3 are displayed in bold.

3 Clustering Simplified German Texts

The aim of the work underlying this thesis was to automatically analyse existing simplified texts to empirically search for evidence of multiple complexity levels in simplified German. Figure 3 illustrates the steps leading to fulfil this aim that are described in this chapter. I firstly introduce the resources I used for this project, summarising the path from text collection to the creation of the datasets. Then, I enumerate a series of experiments performed to discover groups of similar texts. For each experiment, I delineate the process of feature extraction and selection, the choice of the machine learning algorithm, the results and their evaluation.
Unlike the studies introduced in Chapter 2, I did not classify the texts with pre-defined levels and labels (cf. CEFR in Section 1.1), because this approach would presume existing and pre-verified complexity levels of simplified German. Rather, I adopted unsupervised methods, such as clustering algorithms, which partition given sets of texts based on similar or dissimilar characteristics. In doing so, I integrated and empirically analysed the features described in Section 2.3.1.

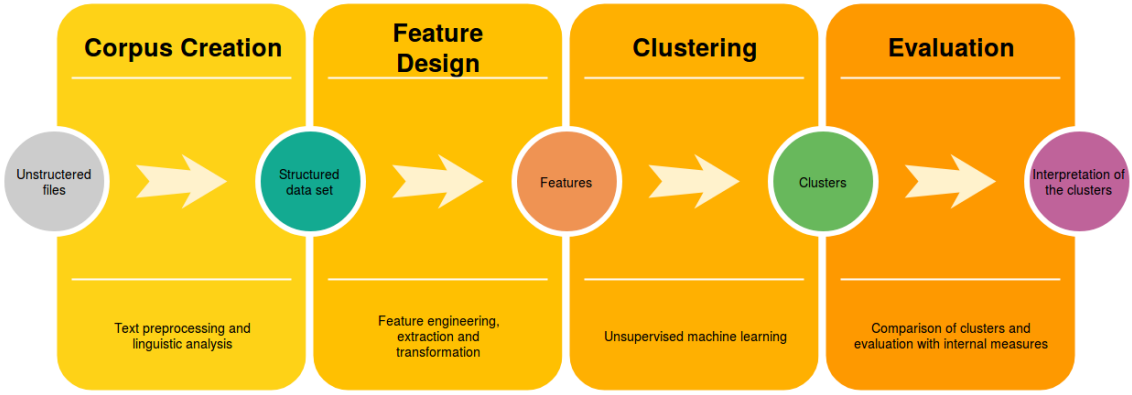


Figure 3: Pipeline of the text clustering analysis.

3.1 Data

The main data used in my analysis consists of a corpus of simplified German (Section 3.1.1) that was created in a project undertaken at the Institute of Computational Linguistics, University of Zurich.[1]
A second dataset was created in the context of this thesis (Section 3.1.2). This dataset includes texts classified at two different levels of simplified German provided by *capito* (cf. Section 1.1) for research purposes. The experiments with this dataset are not meant to be a validation of my clustering method, but rather a replication of the experiment settings using different data in simplified German.

3.1.1 Corpus of Simplified German

Klaper et al. [2013] compiled the first German/simplified German parallel corpus, which consisted of 256 texts amounting to approximately 7,000 sentences and 70,000 tokens. This corpus [Klaper et al., 2013] was extended in the following ways to be suitable for use in both automatic readability assessment and text simplification of German [Battisti and Ebling, submitted]:

- The corpus contains more parallel data.
- The corpus additionally contains monolingual-only data (simplified German).
- The corpus contains new information on text structure, typography and images.

The extended corpus by Battisti and Ebling [submitted] served as the basis of the experiments reported in this thesis. Figure 4 summarises the main steps undertaken to create the corpus. The corpus includes webpages and PDFs, which are of different text types and were composed according to different guidelines (cf. Section 1.1) by various authors. In the German, Austrian and Swiss web domains, Battisti and Ebling [submitted] manually identified 92 publicly accessible websites offering materials in simplified German, including websites of governments, specialised institutions, translation agencies, and non-profit organisations. Then the authors downloaded the original source files from the identified websites between September 2018 and January 2019. For the webpages, the authors manually checked each static dump to verify their content and language and, subsequently, removed HTML markup using the Beautiful Soup library for Python[2]. For the PDF files, they made

[1] Further details about the corpus are described in Battisti and Ebling [submitted].
[2] `https://pypi.org/project/beautifulsoup4/` (last accessed: 28 April 2019)

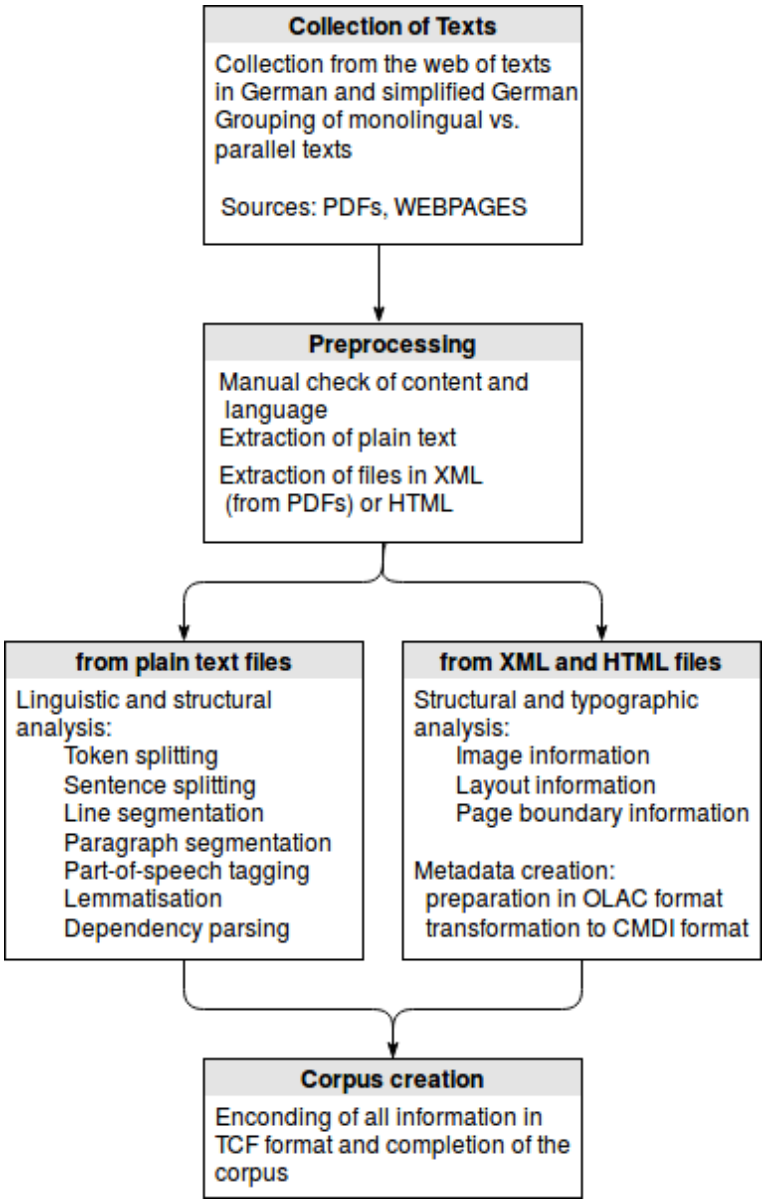


Figure 4: Pipeline of the corpus creation.

use of the PDFlib Text and Image Extraction Toolkit[3] (TET) to extract the content as plain text and information on text structure, typography and images in an XML format, called TETML[4].

The content of each file was analysed automatically and the results were directly converted into TCF documents. TCF, namely *Text Corpus Format*, is an XML data exchange format that was developed within the WebLicht architecture [Hin-

[3] `https://www.pdflib.com/` (last accessed: 28 April 2019)

[4] XML-based format adopted by the PDFlib Text and Image Extraction Toolkit (TET) library.

richs et al., 2010] to support the incremental enrichment of linguistic annotations in a stand-off markup. The structure of each TCF document can be divided into two macro-sections, each consisting of various micro-sections. The macro-sections correspond to the metadata, or primary data, and the text corpus element, or secondary data.

The micro-sections of the metadata store information about the documents, such as title, contributors, unique identifier and date. If specified, the authors also included information about the CEFR language level, the presence of images and the affiliation with a related resource, which could be the prototypical or derivative work in original or simplified German. The metadata was firstly mapped to the Open Language Archives Community (OLAC) Standard[5], which is based on the Dublin Core Metadata Element Set[6], and then converted into the Component MetaData Infrastructure (CMDI)[7], the metadata standard of CLARIN [Hinrichs and Krauwer, 2014] adopted for the TCF format.

The text corpus macro-section is organised into ten different micro-sections or layers, each containing linguistic annotation at various levels of analysis: text, tokens, sentences, the structure of the document (paragraph and line segmentation), lemmas, parts of speech (POS), morphological annotation, dependency parsing, images and fonts. The linguistic analysis for the layers was conducted using ParZu, the Zurich hybrid dependency parser for German that combines hand-written rules with a probabilistic disambiguation system [Sennrich et al., 2009, 2013]. ParZu performs tokenisation and sentence segmentation using the Natural Language Toolkit (NLTK) [Bird and Loper, 2002]. Differently from ParZu's standard configuration, Battisti and Ebling [submitted] adopted the TreeTagger [Schmid, 1994] for POS tagging and Zmorge [Sennrich and Kunz, 2014] for morphological analysis.

Image and font layers are not catalogued in the TCF guidelines available online.[8] Battisti and Ebling [submitted] designed these two layers to preserve significant layout and structural aspects of texts, because they presume these features could be predictive in an automatic readability assessment context, such as the analysis presented in this thesis. As explained in Section 2.3.2, images, fonts and textual formatting devices convey relevant information, which should be considered while examining simplified documents.

[5] `http://www.language-archives.org/OLAC/olacms.html` (last accessed: 11 April 2019)

[6] `http://dublincore.org/` (last accessed: 11 April 2019)

[7] `https://www.clarin.eu/faq/how-can-i-convert-my-dc-or-olac-records-cmdi` (last accessed: 11 April 2019)

[8] `https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format` (last accessed: 11 April 2019)

Information about images and typographic information (e.g., boldface, italics, underline, etc.) was selected from HTML and TETML files and subsequently saved in specific layers. The image layer subsumes pictures, logos as well as drawings included in the resource document. Images were treated separately because of their numerous properties, which might be useful to consider for more fine-grained analyses on image-text relationships as in Bosma [2005]; Cha et al. [2019]; Specia et al. [2016].
Typographic information, such as information on formatting devices, was added to the layer representing the structure of the document, in which a token sequence was annotated as belonging to some text structure element, such as page, paragraph, line and heading. In this way, the relationship between a token sequence and structural element is immediately apparent. The font layer was designed to preserve detailed information about the font(s) inherent in resource files and referenced in the tokens layer. Appendix A shows an example of a TCF file.

Overall, the corpus consists of 6,217 texts. About 95% of the documents were originally in HTML format. Files in PDF format amounted to 272 texts. The tokens for websites and PDF files are shown in Table 4. While the number of websites is clearly larger than the number of PDF files, the number of tokens of the first is almost equal to the latter, indicating the PDF documents are substantially longer than the HTML documents. This is also demonstrated through the average number of sentences per document, which is 320 for the PDFs and 20.81 for the HTML files.

	PDF	HTML
Number of documents	272	5,945
Number of sentences	87,045	123,971
Number of tokens	1,088,119	1,422,849
Avg. no. of sentences per document	320	20.81
Avg. no. of words per sentence	12.5	11.47
Vocabulary size	168,261	546,341

Table 4: Corpus profile: PDF vs. HTML.

Table 5 shows the complete corpus profile. The monolingual-only texts contain fewer sentences on average than the simplified German side of the parallel data (31.64 vs. 55.75). This tendency is further shown in the average length of the texts. The monolingual-only texts contain on average 350.8 tokens, while the corresponding number for the monolingual texts in the parallel section is 651.8. However, the average sentence length is almost equal (ca. 11 words). It can be seen that the monolingual texts are shorter than the simplified version in the parallel texts. This

length discrepancy might be due to a wide range of reasons, such as the topic of the texts and the application of different writing methods and rules. The monolingual parallel corpus section accounts for a minor portion of the corpus, that is, only 378 texts (378 for German and 378 for simplified German) against 5,461 monolingual-only texts. The simplified texts in the monolingual parallel section have clearly undergone some sort of simplification process. As Table 5 shows, they contain on average more sentences, fewer tokens and a reduced vocabulary compared to their original counterparts. Figure 5 shows an example of a text in German alongside with its simplified version.

	German	**Simplified German**
Monolingual		
Number of documents		5,461
Number of sentences		172,773
Number of tokens		1,916,045
Avg. no. of sentences per document		31.64
Avg. no. of tokens per sentence		11.09
Parallel		
Number of documents	378	378
Number of sentences	17,121	21,072
Number of tokens	347,941	246,405
Avg. no. of sentences per document	45.29	55.75
Avg. no. of tokens per sentence	20.32	11.69
Vocabulary size	33,384	16,352
Total of monolingual and parallel texts		
Number of documents		6,217
Number of sentences		210,966
Number of tokens		2,510,391
Avg. no. of sentences per document		33.93
Avg. no. of words per sentence		11.90
Vocabulary size		614,564

Table 5: Complete corpus profile Battisti and Ebling [submitted].

Eine Krankheit oder Behinderungen haben oft große Auswirkungen auf den Arbeitsalltag. Manchmal sind diese so stark, dass Beschäftigte ihren Beruf nicht mehr ausüben können. In diesen Fällen ist eine berufliche Neuorientierung notwendig.

Die Entscheidung für eine andere berufliche Tätigkeit fällt erfahrungsgemäß nicht leicht. Damit Sie leichter eine Entscheidung treffen, können Sie verschiedene Berufstätigkeiten und Arbeitsplätze ausprobieren.

Zum Angebot im Rahmen der beruflichen Neuorientierung zählen insbesondere:

- Berufsvorbereitung einschließlich einer wegen der Behinderung erforderlichen Grundausbildung,
- die individuelle betriebliche Qualifizierung im Rahmen Unterstützter Beschäftigung,
- die berufliche Anpassung und Weiterbildung, auch soweit die Leistungen einen zur Teilnahme erforderlichen schulischen Abschluss einschließen,
- die berufliche Ausbildung, auch soweit die Leistungen in einem zeitlich nicht überwiegenden Abschnitt schulisch durchgeführt werden sowie
- die Förderung der Aufnahme einer selbstständigen Tätigkeit durch die Rehabilitationsträger.

Diese Leistungen werden durch Berufsbildungswerke, Berufsförderungswerke und vergleichbare Einrichtungen der beruflichen Rehabilitation ausgeführt, wenn Art oder Schwere der Behinderung oder die Sicherung des Erfolges die besonderen Hilfen dieser Einrichtungen erforderlich machen.

(a) Standard German

Manchmal muss man einen neuen Beruf lernen.

Zum Beispiel:

- Wenn man krank ist.
- Wenn eine Behinderung schlimmer geworden ist.

Und man seine Arbeit nicht mehr gut machen kann.

Dann muss man darüber nachdenken:

- Welche andere Arbeit macht mir Spaß?
- Was kann ich gut?
- Was möchte ich arbeiten?

Das bedeutet:
Man muss über einen neuen Beruf nachdenken.

(b) Simplified German

Figure 5: Example of a text in German (a) and simplified German (b) from `http://einfach-teilhaben.de/` (last accessed: 30 May 2019).

52 websites come with a language certificate that validates the language level according to a standard, such as the *capito Gütesiegel*[9], while 3,234 websites are *informally* classified as A1.

[9] `https://www.capito.eu/de/Ueber_uns/capito_Qualitaets-Standard/Guetesiegel_Qualitaets-Standard/` (last accessed: 11 April 2019)

Regarding the text genre, all texts are in prose, covering different domains and subjects. The genres can be grouped into larger text categories, such as instructions, news articles, blogs and short stories. A preliminary topic modelling analysis, which I performed to explore the corpus, has confirmed a tendency to social subjects. The topics range from politics (e.g., introduction to the voting system) to health (e.g., instruction for specific medical treatments) and culture (e.g., description of art museums). A brief summary of the topic modelling analysis and the resulting topics are included in Appendix B.

3.1.2 "TopEasy News" Text Collection

Since January 2017, the Austrian Federal Ministry of Labour, Social Affairs, Health and Consumer Protection has been promoting a project of the Austrian Press Agency (APA) entitled "TopEasy News".[10] In this project, news articles covering various topics, such as politics, economics, culture and sport, are translated into simplified German at levels A2 and B1 by *capito* (cf. Section 1.1).

The text collection created for this analysis contains news articles written between August 2018 and March 2019, directly downloaded from the APA website at the beginning of 2019. Each document underwent the same preprocessing pipeline (cf. Figure 4) designed for the analysis of webpages in the corpus presented in Section 3.1.1. To summarise, the main content was extracted by removing HTML markup and boilerplate and then analysed with ParZu [Sennrich et al., 2009, 2013]. The output was saved in TCF format including information on images and typography. For the sake of this thesis, metadata was not created.

The resulting collection includes 1,331 news articles distributed according to three categories, namely original German (OG), B1 language level and A2 language level (cf. Table 6). Not all texts are parallel. For instance, 41 documents at level A2 and 171 documents at level B1 do not have original counterparts. Yet all texts at level A2 have a corresponding translation at level B1. The uneven distribution of the documents in the various categories is not relevant for this thesis, because I only analyse texts in simplified German.

Table 6 shows the complete profile of the text collection. While the number of OG documents is smaller than the number of B1 texts, the number of sentences in the first is almost double the amount of sentences in the latter. This implies the original texts are considerably longer than the B1 translations. The number of tokens for

[10] `https://science.apa.at/site/home/kooperation.html?marsname=/Lines/Science/Koop/topeasy` (last accessed: 11 April 2019)

	OG	B1	A2
Number of documents	428	594	309
Number of sentences	10,474	5,560	3,197
Number of tokens	201,536	71,812	35,033
Avg. no. of sentences per document	24.47	9.36	10.34
Avg. no. of tokens per sentence	19.24	12.91	10.95

Table 6: Profile of "TopEasy News" text collection based on language level.

each category substantially decreases from OG to A2 texts. A2 texts present on average shorter sentences than the original texts (average number of words per sentence: 10.95 vs. 19.24) and their B1 counterpart (12.91). B1 texts on average contain fewer sentences than the A2 texts (average document length in sentences: 9.36 vs. 10.34). A manual verification is needed to identify the cause of this value. It may be a result of insertion of lexical explanation during A2 translation or an incorrect sentence splitting during the preprocessing step.

3.2 Features

In cluster analysis, the selection of variables, or features, has the potential to impact on the performance of the model and change the outcome of the analysis. The features should account for those domain aspects that are pertinent to the research questions [Moisl, 2015]. Therefore, I designed features that attempt to define all relevant factors describing simplified language, especially German. The process of designing, extracting and transforming the variables from unstructured texts is explained in this section.

3.2.1 Feature Design

I designed and extracted text-, sentence- and token-level features that have proven to be distinctive in classification of simplified texts based on the studies presented in Section 2.3.1. In addition, I implemented the features illustrated in Section 2.3.2. To the best of my knowledge, these features have never been included in readability assessment and classification studies yet. The complete feature set contains 115 features grouped into surface features, such as LIX score and character-level variables,

and deeper linguistic features based on lexicon, morphology and syntax of simplified German. Note that I did not apply some common preprocessing steps ahead of feature extraction, such as stopword removal and stemming, even if they are widely adopted in NLP. Stopwords can be psychologically informative [Naigles et al., 2016; Chung and Pennebaker, 2007], while different morphological inflections (over which stemming is designed to abstract) are known to be acquired at different stages of language learning. The table in Appendix C gives an overview of the complete list of features implemented for this analysis.

Surface Features

In this group, I incorporated readability, character-level and layout features. I took advantage of the structural and typographic information included in the corpus of German/simplified German (cf. Section 3.1.1) to explore the features listed in Section 2.3.2. I integrated counts of compounds with the special characters *Mediopunkt* and hyphen. Then, I counted the number of images, paragraphs, lines and the number of words displaying a specific font type (e.g., boldface). Additionally, I implemented the concept of "one-sentence-per-line", which is fundamental to simplified language.

Deeper Linguistic Features

Lexical and Semantic Features: To guarantee the independence of the word lists from the dataset, I used a frequency list created on the basis of the German Reference Corpus (DeReKo, [Institut für Deutsche Sprache, 2014; Lüngen, 2017]). Frequencies were calculated by adding the POS probabilities provided by the TreeTagger [Schmid, 1994]. From this list that includes 100,000 entries, I considered only the top 500 (simple) and the bottom 1,000 (difficult) entries. I exploited two further frequency lists based on the language learning book "Profile Deutsch" [Glaboniat et al., 2005], because they specify the receptive linguistic complexity level for each entry word (approximately 6,966 lexical entries). One list consists of common words in everyday life; the second list enumerates technical and professional words.

Morphological, Morpho-syntactic and Syntactic Features: The set of morphological, morpho-syntactic and syntactic features ranges from identification of impersonal constructions to word and sentence order. Following pronoun extraction, I separately identified and treated first- and second-person pronouns. These two types of personal pronouns are easier to understand than a third-person pronoun by the target readers, with the exception of persons affected by Autism Spectrum Disorder (ASD) [Naigles et al., 2016].[11] I integrated the compound ratio, namely

[11] Persons, in particular children, affected by ASD tend to use the second-person pronoun *you* instead of the first-person pronoun while referring to themselves [Naigles et al., 2016].

the ratio of compound nouns to all nouns [Hancke et al., 2012]. Finally, I explored the comparison between genitive modifiers and the substitute form *von*+dative.

3.2.2 Feature Transformation

For each document, I extracted all features and stored them in a file as comma-separated values. Instead of taking the absolute counts, I normalised them by dividing the values by the length of each document expressed in tokens. Normalisation is necessary when the features are extracted from documents with a significant variation in length, as the analysed resources in this thesis. However, the kind of normalisation I adopted is ineffective when the documents are too short to provide reliable probability estimates, affecting the validity of the clustering [Moisl, 2015]. To avoid this issue, I implemented a filtering function that deleted short samples, in this case documents with fewer than 20 tokens.

A second normalisation process was executed after the data matrix creation but prior to cluster analysis. I applied the `StandardScaler()` function in `scikit-learn`[12] to centralise the features, namely to ensure that for each feature the mean was 0 and the variance 1. This normalisation is fundamental if the data matrix undergoes any dimensionality reduction process, such as Principal Component Analysis (PCA), which aims at reducing n-dimensional vectors as much as possible while diminishing the loss of information. Therefore, I performed dimensionality reduction to have a 2-dimensional condensed representation of the vectors and avoid the "curse of dimensionality" problem [Raschka, 2015], according to which the higher the dimensionality, the more difficult it becomes to define the shape of the cluster.

To summarise, normalisation, standardisation and dimensionality reduction were applied to mitigate the risk of values that would compromise the quality of the analysis. In the first and second experiment (Sections 3.4 and 3.5) and in the discussion (Chapter 4), I address the problem of outliers in the context of this thesis.

3.2.3 Feature Engineering

For my experiments, I considered different combinations of the implemented features introduced in the previous section (cf. Section 3.2.1). The first set included all of the extracted features, namely 115 variables. I subdivided the complete feature set into

[12] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html` (last accessed: 12 April 2019)

subgroups representing linguistic categories as presented in Sections 2.3.1 and 3.2.1. With these subsets, I aimed at comparing the performance of the feature groups in isolation as well as in combination with each other. Table 7 shows the five feature groups that focus on different linguistic and structural aspects of documents.

Subset	Features	Number	Reduction
1	All	115	x
2	Surface only	26	x
3	Deeper only	89	x
4	Lexical and semantic only	17	x
5	Morphological and syntactic only	72	x

Table 7: Subsets of feature combinations; x: application of dimensionality reduction.

3.3 Evaluation Metrics

One of the drawbacks of unsupervised machine learning algorithms is the evaluation process, since the data is unlabelled. Focusing on clustering algorithms, there are also no definitive standards or criteria for selecting the ideal number of clusters.
Before presenting the experiments and their results, I introduce three metrics I considered to choose the optimal number of clusters (k) and to preliminarily evaluate clustering results without requiring a ground truth.
Silhouette coefficient, Calinski-Harabasz index and elbow method are three validation metrics that are generally used to tune the hyperparameters of a model. They can test the final performance of the algorithm, if there is no development set. The proper way to use these criteria is to compare the clustering solutions obtained on the same data by running one or more algorithms with different hyperparameters. The actual application and results of these metrics are illustrated in each designated section (cf. Sections 3.4.2, 3.5.2 and 3.6.2).

The silhouette coefficient is a measure of how compact and well-separated clusters are. For a set of samples, the silhouette score is computed by averaging the silhouette coefficient for each sample, which is calculated as "the difference between the average intracluster distance and the mean nearest-cluster distance for each sample, normalised by the maximum value" [Bengfort et al., 2018, 178]. The score ranges from 1 to -1, where 1 represents highly dense clusters, -1 incorrect clustering and a value around 0 is evidence of overlapping clusters. For example, vertically thick but low scoring clusters suggest that the number of clusters (k) should be higher.

The Calinski-Harabasz index is based on the concept of dense and well-separated clusters and defined as the ratio between the intercluster and the intracluster dispersion. The best result is given by a high intercluster dispersion, which symbolises well-separated clusters, and a low intracluster dispersion, which is evidence of dense clusters.

The elbow method visualises multiple clustering models with different values for k. The elbow in the plotted line corresponds to the best value of k. If data is not well clustered, this method may lead to a smooth or jagged curve that might suggest, inter alia, that the clustering algorithm should be reconsidered. This method accepts as input value both the silhouette score and the Calinski-Harabasz index.

3.4 Experiment 1: Hierarchical Clustering

With this experiment, I aimed at finding the best parameter function available in agglomerative hierarchical clustering for my research questions. With the best experimental settings, I explored the dataset with different feature combinations (cf. Section 3.2.3) and data subsets in order to look for evidence of multiple complexity levels. The clustering results are investigated and discussed in Chapter 4.

Theoretical Background

Agglomerative hierarchical clustering is a general-purpose unsupervised machine learning method, which successively merges items into clusters based on their similarity with one another. At the beginning, each item is considered as a single cluster. Then, the algorithm merges pairs of clusters based on a predefined similarity or distance metric and a linkage criterion to determine the degree of association between clusters. The first parameter, the similarity or dissimilarity metric, is necessary to calculate the distance matrix between two documents or feature vectors, which serves as input for the algorithm. Euclidean distance (aka L2 norm) and cosine similarity are the most used metrics for tasks related to document classification and clustering. The second parameter, the linkage criterion, determines the similarity between two clusters. The standard agglomerative hierarchical algorithms include single, complete and average linkage criteria. With single linkage, the clusters with the smallest distance between items are merged, disregarding distant data points. The complete linkage criterion estimates the similarity of two clusters comparing the most distant, or dissimilar, data points in the clusters. The average linkage criterion works as a compromise between single and complete linkages. It merges cluster pairs based on the minimum average similarity between all pairs of items in the two clusters.

A dendrogram, namely a tree diagram, visualises the hierarchical relationships between clusters. It enables to determine the number of clusters by drawing a cut-off line at a predetermined value and counting the number of lines that this cut-off line intersects (cf. Figure 6).

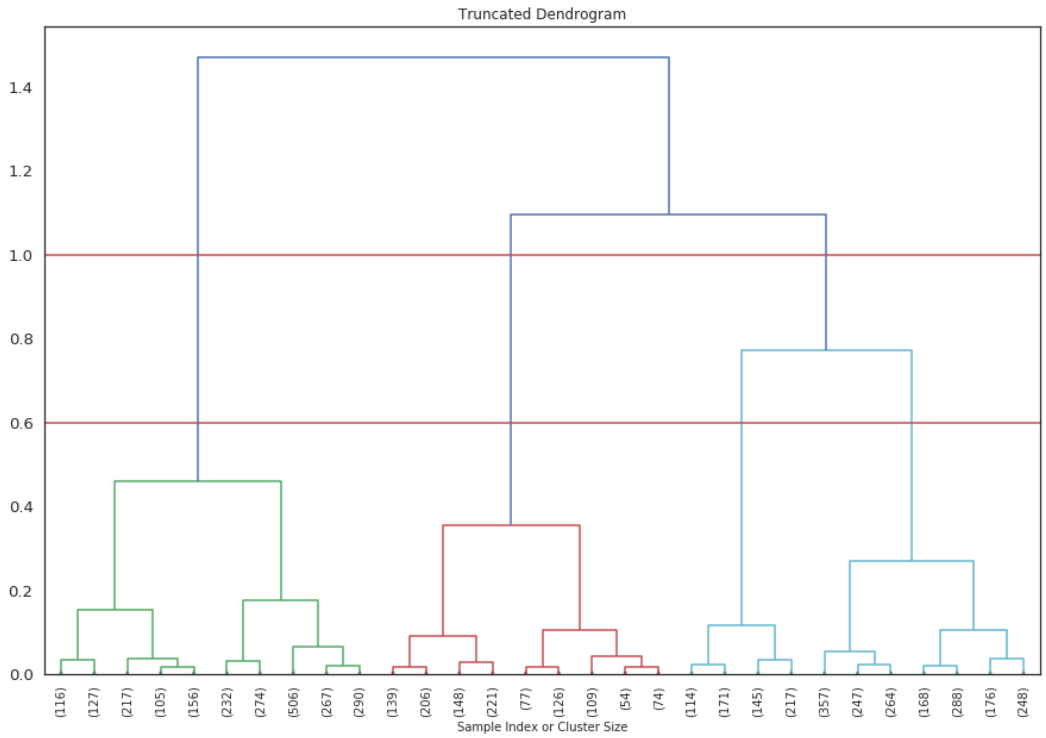


Figure 6: Sample of a truncated tree considering all features.

3.4.1 Method

For the hierarchical clustering, I used the `linkage` function[13] from the `Scipy` library. Since there is no rule of thumb to select the metric and linkage parameters (cf. Section 3.4), I chose the metric-linkage combination that yielded the highest Cophenetic Correlation Coefficient (CCC)[14] [Sokal and Rohlf, 1962] by considering the complete feature set (cf. Subset 1 in Table 7). The CCC is a correlation coefficient between the distance matrix (input of the agglomerative hierarchical clustering) and the cophenetic matrix, which includes the cophenetic distances between each item in the hierarchical clustering.
For each metric-linkage pair, I first calculated the distance matrix using the `pdist`

[13] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage` (last accessed: 14 April 2019)

[14] `https://docs.scipy.org/doc/scipy-0.16.1/reference/generated/scipy.cluster.hierarchy.cophenet.html` (last accessed: 14 April 2019)

function from `spatial.distance` submodule in `Scipy` and called the `linkage` function from `cluster.hierarchy`, which returned the linkage matrix. Then, I calculated the CCC to measure and compare the quality of the output dendrograms. The closer to 1 the cophenetic index is, the better the dendrogram is in preserving and representing the original distances.[15]

Linkage \ **Metric**	Cosine	Euclidean
Single	0.563	0.655
Complete	0.681	0.57
Average	**0.797**	**0.831**

Table 8: Comparison of the CCC results for the data.

Both Euclidean and cosine metrics reached the highest CCC indices with the average linkage (cf. Table 8). Since Euclidean distance is known to perform poorly on high-dimensional data [Sohangir and Wang, 2017] and to be sensitive to normalisation and scaling techniques, I only took into consideration the model based on cosine similarity metric.

The output of the `linkage` function is suitable for creating tree diagrams using the `dendrogram` function[16] from the `Scipy` library. To select the value of the cut-off line(s), I considered the height of the dendrogram vertical bars (i.e., branches). It is commonly known that the greater the difference in height of the branches, the more dissimilar the clusters are [Rogel-Salazar, 2017]. For example, in the dendrogram in Figure 6, drawing a cut-off line at the value 1.0 results in three clusters, while drawing a cut-off line at the value 0.6 gives four more fine-grained clusters. Theoretically, it is possible to draw a line at 0.3 and retrieve 7 clusters, but the difference among them would be smaller as the difference in height of the branches indicates.
In Section 3.3, I described three additional methods commonly used to determine the optimal number of clusters and thus internally evaluate the clustering method. In the next section, I compare the results.

For the agglomerative hierarchical clustering algorithm, I also made use of the `scikit-learn` toolkit[17] to recursively create models and calculate the evaluation

[15]Studies have shown that datasets with outliers have higher cophenetic correlation values than datasets without outliers [Saraçli et al., 2013].

[16]`https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html#scipy.cluster.hierarchy.dendrogram` (last accessed: 14 April 2019)

[17]`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.`

metrics. This ensured the appropriate number of clusters to retrieve. To avoid differences in implementation between `Scipy` and `scikit-learn`, I made sure that the two implementations were equivalent by running some examples side by side and observing that the results were the same, especially for the chosen k.

Considering the selected k, the clusters were retrieved using the `fcluster`[18] function from the `Scipy` library, which forms flat clusters from the dendrogram. This enabled me to gain a more detailed insight into the clusters, the texts included in the clusters and the features characterising these texts (Chapter 4).

I repeated the whole procedure according to the five aforementioned feature subsets. Additionally, I carried out two further subexperiments with this method. In the first subexperiment shown in Section 3.4.2.1, I separated the PDFs from the webpages and repeated the procedure explained above on these two independent subsets. This exploration was motivated by a desire to understand whether source format and quality had an influence on the clustering algorithm and how features were distributed among qualitatively different types of resources. The intuition behind this is that the quality of the PDF texts is better than that of the webpages, because PDF files are usually checked by translation providers specialised in simplified German.[19]
In the second subexperiment (Section 3.4.2.2), I applied agglomerative hierarchical clustering after feature agglomeration. While agglomerative hierarchical clustering recursively merges samples to create clusters, feature agglomeration groups together features that behave similarly. This clustering approach can be used to reduce the dimensionality of the feature matrix instead of PCA (cf. Section 3.2.3) and allows for an easy interpretation of the results. Therefore, after feature extraction, I applied feature agglomeration using the `sklearn.cluster.FeatureAgglomeration` submodule[20] from the `scikit-learn` library. The purpose of this experiment was to examine whether feature agglomeration of sparse features could affect and improve the clustering performance.

`AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering` (last accessed: 14 April 2019)

[18] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html#scipy.cluster.hierarchy.fcluster` (last accessed: 14 April 2019)

[19] Almost each PDF file includes a reference to a specialised proofreader. This information is generally not specified in the webpages, with some exceptions.

[20] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.FeatureAgglomeration.html` (last accessed: 14 April 2019)

3.4.2 Results

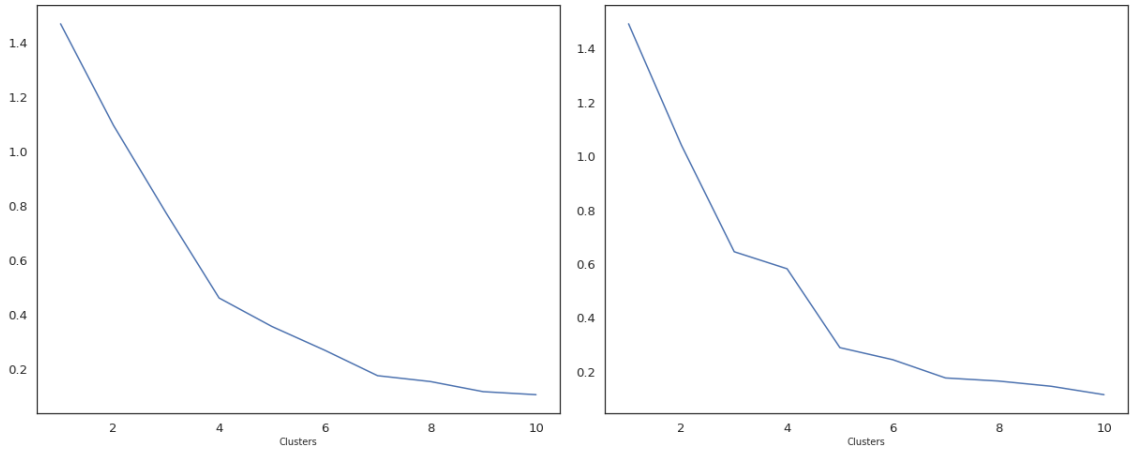
3.4.2.1 Without Feature Agglomeration

Complete Dataset

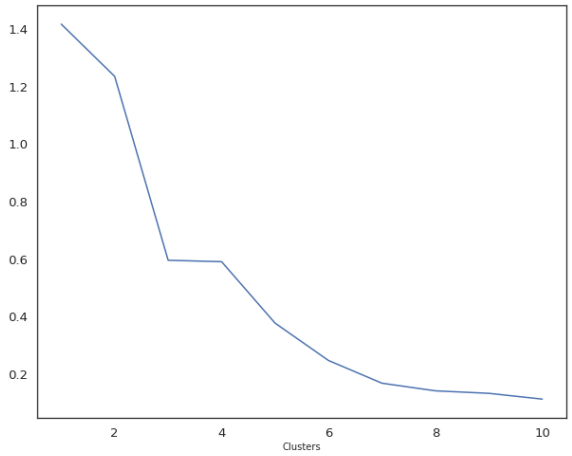
Table 9 summarises the results of the evaluation metrics (cf. Section 3.3), computed by running the agglomerative hierarchical algorithm with multiple cluster sizes k to compare how this parameter k affects the clustering results. All feature subsets achieved similar silhouette and Calinski-Harabasz scores. In general, the silhouette coefficients reached higher values as the number of k increased, indicating better cluster quality. Subset 1 showed, however, high silhouettes scores when $k \geq 7$. In three subsets (2, 4 and 5), the algorithm set to $k=3$ performed better than the other algorithm settings, suggesting that $k=3$ is the optimal number of clusters in this dataset. The elbow visualisation confirmed these results. Figure 7 illustrates the results obtained with the elbow method for the Subsets 1, 2 and 5. For the first subset, no obvious "elbow" is visible, hence also from this plot it is difficult to choose the optimal size of k. The plots of the elbow method for the second and fifth subsets support the results in Table 9. Both plots show an "elbow" corresponding to $k=3$. All this considered, I analysed the features included in the subsets according to a three-cluster solution. Figure 8 shows an excerpt of a sample text for each cluster resulting from the analysis with the second subset. The average feature analysis showed that Cluster 1 consisted of structured and short texts following the rule "one-sentence-per-line" and generally including only one image (cf. left text in Figure 8a). Conversely, Clusters 2 and 3 contained long and less structured texts, with numerous or absent images and words in bold type (cf. Figure 8a and b).

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
2	0.642	**3143.9**	0.662	2306.1	0.605	**3638.4**	0.538	1729.3	0.528	3353.4
3	0.617	2571.1	**0.685**	**2396.1**	0.585	3133.2	**0.649**	**2062.7**	**0.657**	**3605.8**
4	0.601	2070.8	0.647	1838.2	**0.63**	2951.8	0.624	1876.9	0.591	2915.2
5	0.622	1805.8	0.604	1628.6	0.580	2519.2	0.526	1536.4	0.579	2684.8
6	0.596	1572.5	0.668	1476.6	0.578	2361.3	0.467	1292.1	0.648	2556.9
7	**0.650**	1421.6	0.671	1314.4	0.575	2160.1	0.556	1181.9	0.646	2301.4
8	0.635	1291.9	0.622	1161.8	0.602	1971.5	0.612	1109.8	0.619	2078.3
9	0.622	1154.5	0.586	1022.5	0.620	1780.7	0.618	1015.4	0.580	1884

Table 9: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH). The bold entries represent the best results.



(a) Subsets 1 and 2



(b) Subset 5

Figure 7: Plots of the elbow method for feature subsets 1, 2 and 5.

Returning to the first subset, the evaluation metrics did not converge in a precise k making it difficult to decide at this stage, which k value fit best in my analysis. In this case, it is recommended to visualise the silhouette profile or the hierarchical tree, as I did by plotting the dendrogram in Figure 6 [Raschka, 2015]. An inspection of the clusters in this figure reported the red cluster to contain only texts from webpages: 757 from the online dictionary Hurraki website and 111 from various non-profit associations, whose texts were written in an extremely simple way. Focusing on the features, the red cluster lacked semicolons, indirect speech constructions, subordinate clauses, words in italics, subjunctive verbs and answering particles. Interestingly, the cyan cluster contained the highest value of words in italics (on average 0.000142) and did not include any present-participle construction. No features were missing in the green clusters.

Karim betritt den Platz.
Er schießt den Ball.
Er rennt.

Er ist gut.
Er ist schnell.
Er ist stark.

Dann bleibt er stehen.

Die Kinder sehen ihn an.
Und die Eltern auch.

Karim lächelt.
Die Eltern lächeln ihn an.

- Beratungs-Stellen, die bundes-weit arbeiten.
  Diese Beratungs-Stellen arbeiten
  in ganz Deutschland.

- Beratungs-Stellen für die Bundes-Länder.
  Diese Beratungs-Stellen arbeiten
  immer für 1 Bundes-Land.
  Zum Beispiel:
  Für Sachsen-Anhalt.
  Oder für Hessen.

- Beratungs-Stellen in den Städten und Gemeinden.
  Diese Beratungs-Stellen arbeiten für die Menschen,
  die dort leben.
  Vielleicht gibt es in Ihrer Nähe
  auch eine Beratungs-Stelle für das Persönliche Geld.

(a) Cluster 1 and 2

Wenn sich ein neuer Betreuer oder eine neue Betreuerin bewirbt, dann gibt es ein Bewerbungs-Gespräch.
In diesem Bewerbungs-Gespräch wird auch über das Thema Sexualität gesprochen.

Es wird darüber gesprochen, was der Betreuer oder die Betreuerin alles tun muss.

Es wird auch darüber gesprochen, was der Betreuer oder die Betreuerin **nicht tun muss** oder **nicht tun darf**.

Es wird auch darüber gesprochen, ob es schwierige Situationen geben kann.
Zum Beispiel:
Es gibt Menschen mit Beeinträchtigungen, die nicht wissen, dass sie jemanden sexuell belästigen oder dass sie einen sexuellen Übergriff machen. Der neue Betreuer oder die neue Betreuerin erfährt, wenn es solche Menschen in der Einrichtung gibt. Das erfährt der Betreuer oder die Betreuerin beim Bewerbungs-Gespräch.

Es kann sein, dass ein neuer Betreuer oder eine neue Betreuerin die Wünsche und Bedürfnisse von den Menschen mit Beeinträchtigungen im Bereich Sexualität nicht ernst nehmen möchte. Wenn das so ist, dann kann er oder sie nicht eingestellt werden.

(b) Cluster 3

Figure 8: Example of a snippet of text for each cluster from the three-cluster solution using Subset 2. Texts from `https://einfachebuecher.de`, `http://www.menschzuerst.de`, `http://www.ki-i.at` (last accessed: 30 May 2019).

In summary, I observed that a value of k between 2 and 4 seemed to be a good cluster solution for the whole corpus according to the three evaluation metrics. The visualisation of the dendrograms (cf. Figure 33 in Appendix D) corroborated the results suggested by the evaluation metrics. Inspecting the variables characterising the clusters demonstrated that linguistic differences were evident among groupings of texts (cf. Figure 8). These findings provide a first indication that current texts

are not simplified at a unique complexity level of German and question the role of images in simplified texts (cf. Sections 2.3.2 and 4.1).

Webpages vs. PDFs

Table 10 shows the results of the evaluation metrics computed in the first subexperiment comparing webpages and PDFs. None of the feature subsets reached the highest values in the same iterations where the complete dataset did (cf. Table 9). For example, in Subset 1, the whole corpus achieved the highest silhouette score of 0.65 in line with *k=7*, while webpages reached a score of 0.647 when *k=4* and PDFs obtained a score of 0.724 in line with *k=3*.

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
					Webpages					
2	0.583	2180.3	0.610	2195.5	0.633	**3688.2**	0.629	**2374.3**	0.574	**3311.8**
3	0.638	**2263.1**	0.579	2533.5	0.588	3042.6	**0.632**	2039.3	0.620	2964.3
4	**0.647**	1940.8	0.591	**2695.7**	0.49	2428.2	0.583	1575.4	0.629	2907.5
5	0.604	1651.5	0.631	2277.5	0.567	2270.2	0.571	1430.3	**0.638**	2695.7
6	0.628	1520.4	**0.633**	1933.2	**0.654**	2250.5	0.526	1247.9	0.630	2404.6
					PDFs					
2	0.649	151.5	0.648	103.8	0.596	115.5	0.662	120.6	**0.702**	214.3
3	**0.724**	**187.1**	0.667	91.1	**0.678**	**197.8**	0.645	**195.4**	0.679	**224.4**
4	0.634	159.3	**0.741**	**122.5**	0.64	165.3	0.671	154.2	0.57	199.3
5	0.558	136.4	0.692	106.8	0.572	166.7	**0.70**	127.7	0.64	188.3
6	0.618	128.1	0.691	106.9	0.664	159.1	0.682	126.4	0.676	183.6

Table 10: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) for webpages and PDFs.

Overall, the silhouette scores are not consistent with the Calinski-Harabasz values. This discrepancy is more evident for webpages than for PDFs and could be attributed to the quality of the texts. In addition, within the same website, webpages are often written by different authors, which may not be trained in simplified German in the same way. Moreover, the texts are often not proofread by experts or target readers. For instance, Hurraki (cf. Section 2.3) generally classifies their articles at a level A1 in the CEFR framework.[21] However, my analysis discovered that their entries distributed across various clusters, indicating that they are clearly not at the same simplified level. At a first glance, I could identify the major difference among these clusterings in the type of described entry. On one hand, clusters mostly contained

[21]Information from a personal correspondence via e-mail with the Hurraki creators.

simple words, such as "Fussball" (en: *football*).[22] On the other hand, clusters consisted of split compounds or complex concepts, such as "Video-Überwachung" (en: *video surveillance*)[23] and "Anschläge am 13. November 2015 in Paris" (en: *Attacks on 13 November 2015 in Paris*).[24]

Among the feature subsets, PDFs obtained more uniform values than webpages. According to the values in the table above, the PDF texts can be grouped into four clusters considering the surface features (Subset 2) or into three clusters considering deeper features (Subset 3) and all features (Subset 1). The low Calinski-Harabasz values for PDFs (e.g. 91.1 in Subset 2) revealed that the clusters were not compact, probably due to the small amount of files in this format.

Figure 9 shows an example of a text for each cluster resulting from the analysis for the Subset 1. The feature analysis for the this subset demonstrated that two clusters consisted of texts which had in general properties commonly classified as simple. The first cluster included texts without prepositions governed by the genitive case, words at a B2 perceptive level and with a minimal rate of hyphens connected to a small ratio of compounds (on average: 0.000005 vs. 0.00004 and 0.0001) (cf. left text in Figure 9a). Cluster 2 especially showed a lack of present-participle constructions and some types of unclassified secondary clauses and the highest ratio for negative constructions (on average: 0.53 vs. 0.29 and 0.071) (cf. right text in Figure 9a). Conversely, Cluster 3 displayed properties related not only to simplified language, such as a low compound ratio (0.0001), but also to complexity, such as secondary clauses. In addition, Cluster 3 exhibited a high rate of words in bold type (on average: 0.44 vs. 0.04 and 0.26), whose level of perception may depend on the textual genre and content (cf. Figure 9b).

Due to the contrasting values across feature subsets, I was able to define the cluster size for webpages through a visual inspection of the silhouette profiles. Figure 10 shows the plots of the silhouette for two different k sizes. The coefficients in the two plots are distant from 0, indicating a good cluster solution. Yet the visible difference in width and length of the silhouettes and the values below 0 are indicators of a suboptimal clustering performance.

Figures 11 and 12 show that both webpages and PDFs can be represented by three main clusters with several more fine-grained subclusters.

[22] `https://www.hurraki.de/wiki/Fu%C3%9Fball_(Spiel)` (last accessed: 4 May 2019)

[23] `https://hurraki.de/wiki/Video-%C3%9Cberwachung` (last accessed: 4 May 2019)

[24] `https://hurraki.de/wiki/Anschl%C3%A4ge_am_13._November_2015_in_Paris` (last accessed: 4 May 2019)

(a) Cluster 1 and 2

(b) Cluster 3

Figure 9: Example of a snippet of text for each cluster in the analysis of PDFs. Texts from `https://einfachebuecher.de/`, `www.frauennotruf-muenster.de/`, `https://jrk-bayern.de/` (last accessed: 30 May 2019).

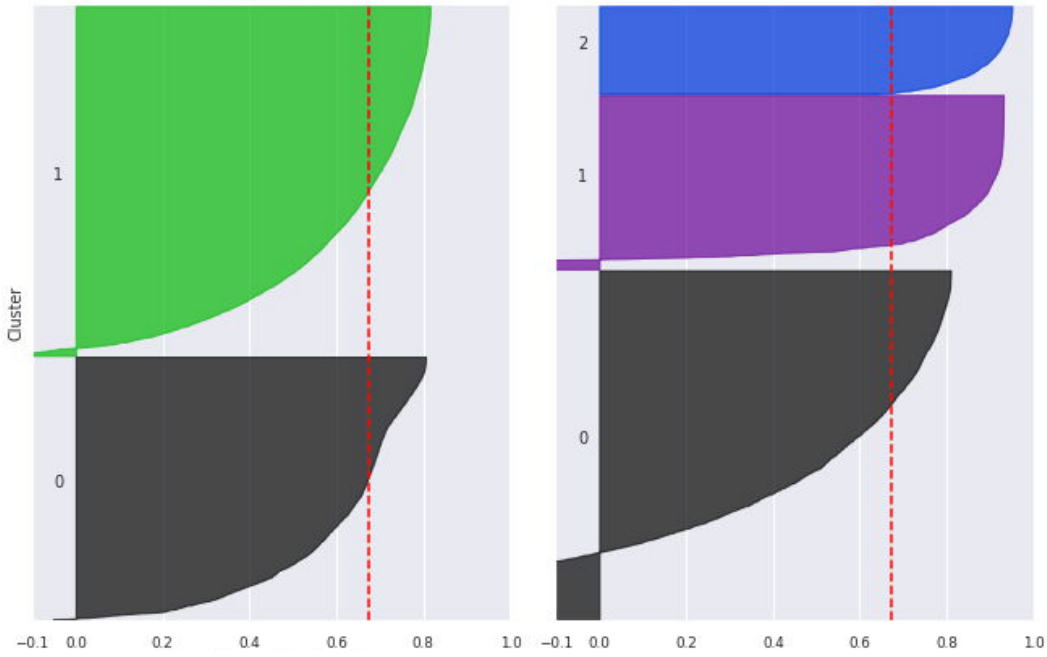


Figure 10: Plot of the silhouette profiles for 2 and 3 clusters considering all features.

To sum up, this visualisation confirms for both text subsets a general tendency of good clustering in line with *k=3*, which is similar to the solution suggested by the analysis of the complete dataset. The exploration of the results proved again that the linguistic differences are observable among groupings of texts in both subsets (cf. Figure 9). In addition, for webpages, the inspection of the files demonstrated that the lexical choices led the clustering to a bipartite, simple vs. complex division of the files.

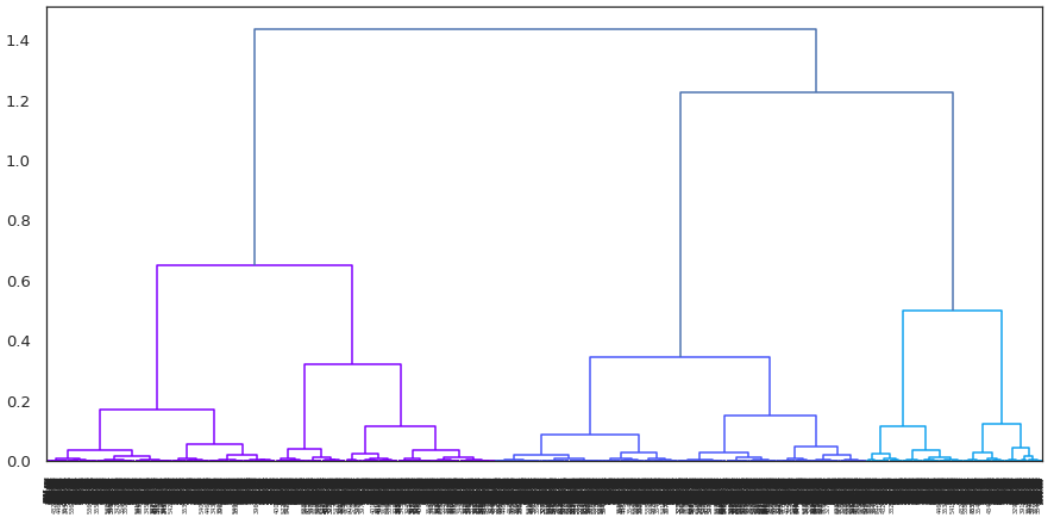


Figure 11: Complete dendrogram of webpages considering Subset 1.

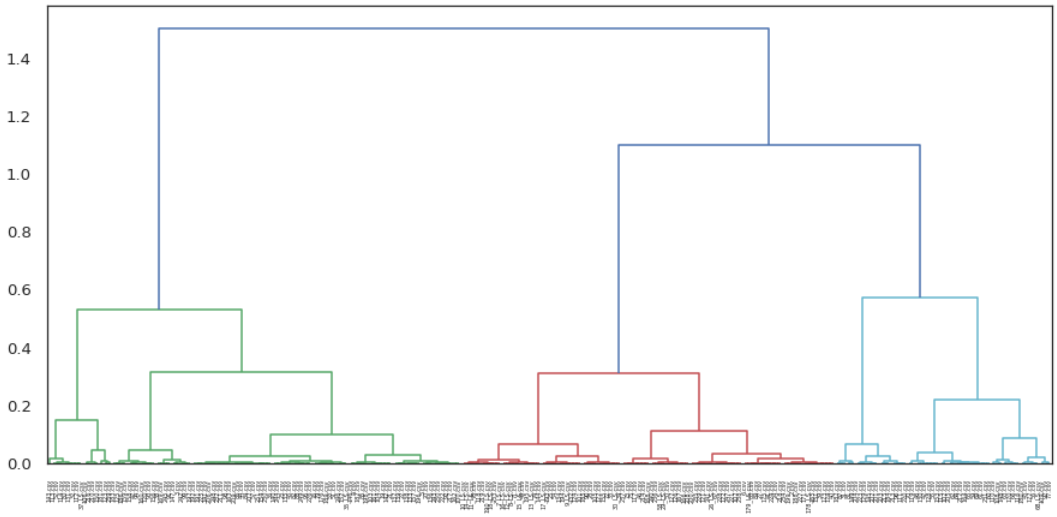


Figure 12: Complete dendrogram of PDFs considering Subset 1.

3.4.2.2 With Feature Agglomeration

The second subexperiment focused on the variables qualifying the samples. The feature agglomeration module in `scikit-learn` requires a predefined number of clusters for the training, which can be set using `n_clusters`. Thus, I applied the algorithm twice by selecting two different values for `n_clusters`: *n_clusters=3*, which is the value for k suggested by the results presented in the previous section (cf. Section 3.4.2), and *n_clusters=5* corresponding to the linguistic categories in the complete feature set.

Complete Dataset

Table 11 summarises the results of the evaluation metrics computed after the feature agglomeration step. Compared to Table 9, silhouette and Calinski-Harabasz metrics showed a consistent tendency for high values across all feature subsets, except for the second. After a closer inspection of the table, I found that the Calinski-Harabasz indices were definitely higher than those in Table 9, indicating more compact and well-separated clusters. On average, the scores of all feature subsets increased by 36% and 23% when the feature space was reduced to 3 and 5 clusters, respectively.[25] In these subsets, an optimal two-cluster distribution of the texts is evident for both algorithm settings. The second feature subset, namely the surface set, obtained lower results than the version without feature agglomeration in that the Calinski-Harabasz score was lower by almost 1000 pt. (cf. Table 9, *k=3*: 2396.1 vs. 1266.3

[25]The percentage is calculated on the iteration on *k=2*, where the feature agglomeration step achieved the highest value.

and 907.3). In total, this subset only comprises 26 sparse features and could hence benefit from a reduction of the space to `n_clusters=2` in order to minimise the noise introduced by sparsity.

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
					n_clusters=3					
2	**0.601**	**3867.1**	0.373	1135.2	**0.675**	**5214.2**	**0.693**	**3593.9**	**0.695**	**5463.2**
3	0.532	2476.2	0.372	1266.3	0.617	3329.5	0.55	1824.8	0.572	3273.9
4	0.456	1698.3	**0.493**	**1417.6**	0.592	2572.7	0.505	1248.9	0.51	2517.8
					n_clusters=5					
2	**0.585**	**3844.2**	0.32	904.5	**0.653**	**5257.8**	**0.52**	**2294.9**	**0.65**	**5145.2**
3	0.45	2138.6	0.31	907.3	0.542	2944.6	0.436	1410.3	0.515	2790.0
4	0.401	1442.7	**0.348**	**982.21**	0.533	2303.5	0.42	1313.1	0.469	2033.2

Table 11: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) after feature agglomeration on all data samples. This table shows only the first three iterations, since the following ones did not obtain relevant results.

For the first subset, the exploration of the features defining the clusters showed a tendency corresponding to the simple/complex dichotomy. I identified the main differences between the two clusters in relation to the following features: nouns, verbs, subjunctive verbs, interrogative clauses, fonts, words in italics and words in bold. In relation to Cluster 2, the first cluster displayed a higher frequency of nouns (0.31 vs. 0.24) and adjectives (0.9 vs. 0.6) linked with a low frequency of verbs (0.13 vs. 0.17). Cluster 2 included also a high rate of images (0.008 vs. 0.004).

In summary, the clustering model benefitted from the feature agglomeration step, which highlighted the discrepancy in feature frequencies across clusters. The plots of the agglomerated features which led to these findings are visualised and discussed in Chapter 4 (Figure 32a).

Webpages vs. PDFs

For both subsets, the feature agglomeration algorithm trained on *n_clusters=3* performed better than the algorithm trained with the parameter *n_clusters=5*. Table 12 summarises the results of agglomerative hierarchical clustering, after grouping the features into three clusters. Compared to Table 10, the feature agglomeration step led to an improvement of 1520 pt. and 46.2 pt. in the Calinski-Harabasz scores of the webpages and PDFs, respectively. This suggests a correlation between the feature sparsity reduction in the data matrix and the creation of more compacted

clusters. Yet for the PDFs, the values of Subset 4 slightly decreased for the silhouette coefficient (0.662 vs. 0.575) and Calinski-Harabasz index (120.6 vs. 117.3) in the first iteration. Besides, Subset 2 shows an inconsistency in the values of the two metrics, which can be explained by the limitations of internal validation measures [Liu et al., 2010] (cf. Section 4.2). It is clear from this table that both subsets generally reached the highest values when texts were grouped into two clusters.

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
Webpages										
2	**0.657**	**4316.2**	**0.708**	**4451.0**	**0.651**	**4541.7**	**0.682**	**3262.9**	**0.682**	**4862.8**
3	0.509	2368.0	0.724	2771.2	0.567	2729.9	0.627	2369.0	0.560	2911.1
4	0.494	1824.4	0.656	1936.5	0.526	2148.4	0.589	1706.4	0.467	2365.4
PDFs										
2	**0.705**	**201.7**	0.587	**138.0**	**0.672**	**207.9**	**0.575**	**117.3**	**0.744**	**271.7**
3	0.645	131.5	0.662	114.5	0.669	157.4	0.545	107.6	0.651	172.9
4	0.51	88.0	**0.667**	115.3	0.573	111.8	0.503	92.3	0.623	131.4

Table 12: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) after feature agglomeration on webpages and PDF files.

Through a visual inspection of the dendrogram, I could closely examine the clustering distribution for the second feature subset and confirm the tendency of the texts to group into two main clusters (cf. Figure 13). The analysis of the surface feature set reported some differences between the two clusters. Among them, the more distinct features were the usage of quotation marks (0.08 vs. 0.001), special characters (0.7 vs. 0.002), digits (0.02 vs. 0.005) and words in bold (0.543 vs. 0.057). The outcome about commas (0.01 vs. 0.02) and exclamation marks (0.0004 vs. 0.0014) supported my assumption of the importance of considering the punctuation marks individually, when dealing with simplified language. Contrary to my expectations, words in italics were more frequent in the cluster defined by the highest rate of boldface than in the other. An opposite distribution of italics and boldface typefaces would justify the claims that italics make the reading comprehension process difficult by changing the form of the characters, while boldface facilitates the process by marking the most relevant expressions in a text (cf. Section 2.3.2).

To sum up, I observed that the clustering models benefitted from the feature agglomeration step also for the webpages and PDFs and grouped the texts into two clusters. A future comparison between the features delineating these clusters and the features in the resulting clusters of the same experiment without feature agglomeration is

necessary to confirm the beneficial effect in the clustering process.

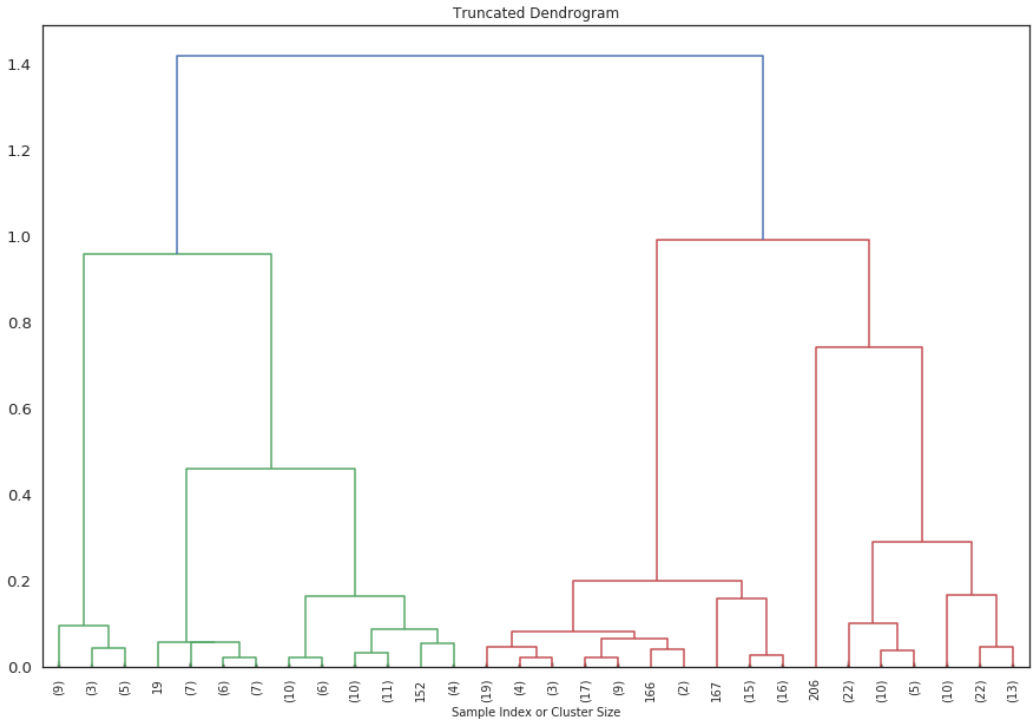


Figure 13: Visualisation of the truncated dendrogram after feature agglomeration considering the second feature subset for the PDF files.

3.5 Experiment 2: K-means

In this experiment, I made use of the k-means algorithm to analyse the data in simplified German. The choice of this algorithm was taken because k-means is one of the most widely used partitioning clustering technique in document clustering research [Aggarwal and Zhai, 2012]. It is easy to implement and computationally more efficient than other clustering algorithms. As in the previous experiment (Section 3.4), I investigated the best parameter function for my research questions and explored the dataset with different feature engineering (Section 3.2.3) and text subsets in order to look for evidence of multiple complexity levels. In Section 3.5.2, I present the results of this analysis and compare them with the results of the agglomerative hierarchical clustering experiment.

Theoretical Background

K-means algorithm requires a set of k centroids, whose number has to be defined prior to clustering analysis. Each centroid is the average of one cluster, around which further similar data points locate. Initial centroids are usually found by randomly

selecting a set of seeds from the dataset or by specifying a value in the initialisation parameter. The initial selection of the set of seeds influences the behaviour on the k-means clustering [Aggarwal and Zhai, 2012, 94].
The algorithm iteratively relocates each data point to its closest centroid to minimise the distortion between the representative centroid and one data point. The iterative process is repeated until the algorithm converges, namely the assignment of data points no longer produces any variation from one iteration to another, or until the change in the distortion evaluation falls below a predefined threshold. In other words, the goal of k-means is the minimisation of the distortion, which is measured by cluster inertia or intracluster sum of squared errors (SSE) [Raschka, 2015]. This measure serves as an estimate of the internal coherence of the clusters.[26] In a high-dimensional space, it tends to fall into the "curse of dimensionality" trap (cf. Section 3.2.2), since it is not normalised and relies on the Euclidean distance. Dimensionality reduction is therefore highly recommended prior to k-means clustering [Raschka, 2015] to reduce the features and speed up the overall computation.

3.5.1 Method

As k-means algorithm, I applied the function available in the `scikit-learn` library. For the initialisation of the seeds, I set the parameter `init=k-means++` to the class `KMeans`. This method might lead to an improvement of the results while placing the initial centroids "far away from each other" [Arthur and Vassilvitskii, 2007; Raschka, 2015]. I set the other parameters of the algorithm to default values.

As in the first experiment (Section 3.4), I recursively created clustering models and calculated the evaluation metrics to determine the optimal number of clusters to retrieve. In this way, I avoided the initial specification of the parameter k or `n_clusters` and determined the best partitioning method for my dataset. Then, I repeated the procedure using the five feature subsets (Section 3.2.3) to analyse the power of linguistic and structural features in isolation and combination. Following the same procedure, I clustered webpages and PDFs separately. Finally, I agglomerated the features for each feature subset to investigate the feature behaviour as well as the overall clustering performance.
Using the `scikit-learn` library, it was possible to retrieve the centroids and the cluster labels, which were mapped to the list of files in order to visualise the clusters.

[26] `https://scikit-learn.org/stable/modules/clustering.html#k-means` (last accessed: 7 May 2019)

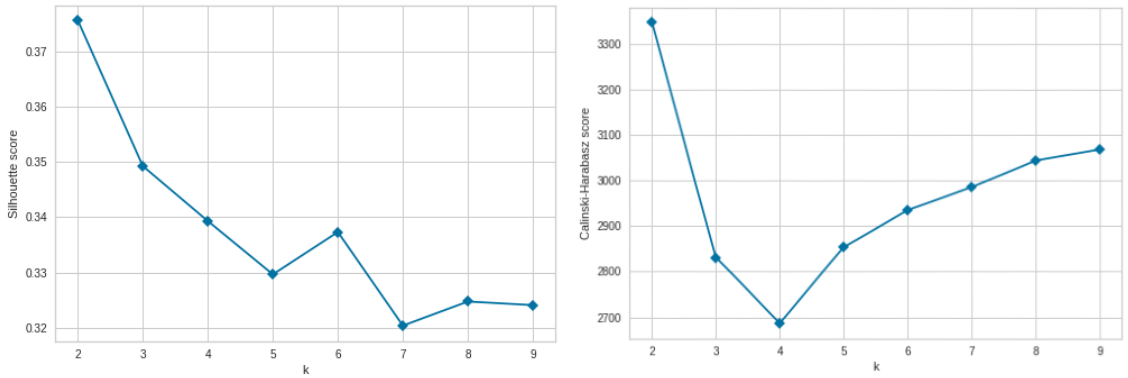
3.5.2 Results

3.5.2.1 Without Feature Agglomeration

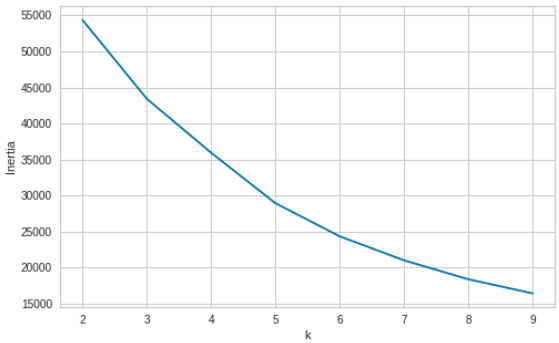
Complete Dataset

As explained in the technical background, k-means includes a randomised parameter, meaning that clusters can vary between different runs on the same input data. The results of this clustering were validated and evaluated not only by combining the three measures described in Section 3.3, but also by visualising the best settings applying PCA, after computing the cosine similarity of the data points.

Figure 14 illustrates the silhouette coefficients, the Calinski-Harabasz scores and the intracluster SSE for the complete feature set (Subset 1). The intracluster SSE almost linearly decreases as the number of clusters increases, making it difficult to determine with certainty the size of k from this plot. On the other hand, the silhouette and Calinski-Harabasz metrics reached the highest value in line with k=*2* and displayed an "elbow" when k corresponded to 5 and 4, respectively.



(a) Silhouette score and Calinski-Harabasz score

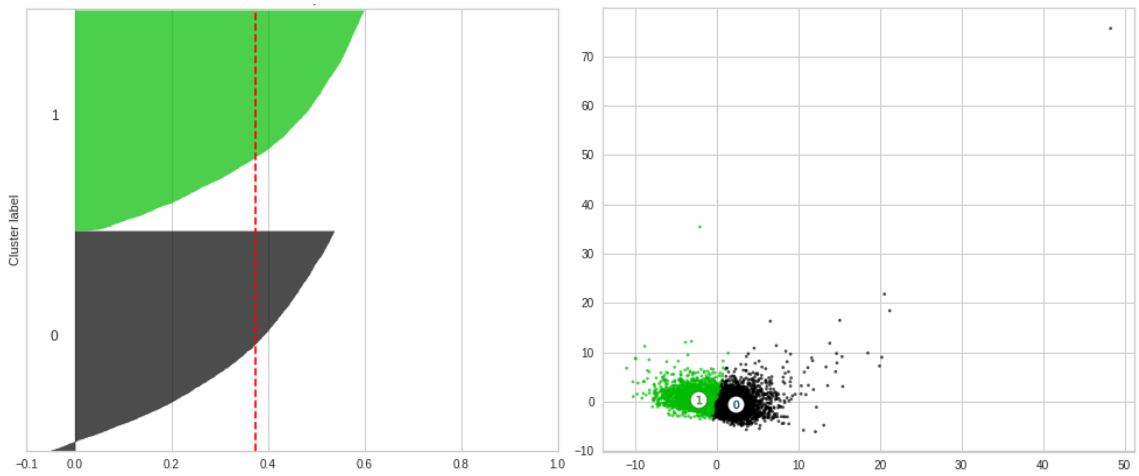


(b) Inertia

Figure 14: Plots of the elbow method for different number of clusters using Subset 1.

As in the experiment using the hierarchical clustering algorithm, the first subset

yielded unclear numerical results (cf. Table 9), so I performed the final evaluation by observing the dendrogram. Similarly, I visualised the silhouette profiles and compared two-dimensional PCA plots for the clustering results. Figure 15 shows the silhouette profile and the scatter plot of the best k for the first feature subset according to k-means clustering. The silhouette profile validated the two-cluster solution, showing a well-shaped form for both coefficients. It is important to notice that, if the dataset had a guaranteed quality, the behaviour of the value of *k=2* would not be optimal. This is confirmed by the outliers that are visible on both plots, namely the values below zero on the left plot and the spread data points on the scatter plot.



(a) Silhouette profile (b) Scatter plot

Figure 15: Visualisation of the best k for the first feature subset.

Turning now to the other feature subsets, the silhouette coefficients and the Calinski-Harabasz indices obtained values in disagreement (cf. Figure 18). With regard to the Subsets 2 and 4, namely surface and lexico-semantic feature sets, the two metrics obtained varying results. For instance, in line with *k=4*, silhouette coefficients reached the highest value, while Calinski-Harabasz indeces achieved the lowest score. Despite the discrepancy, this behaviour accounts for a substantial change in clustering, probably indicating that *k=4* is a good clustering setup. However, considering only the second subset, the visualisation in Figure 16 shows that the clusters are partially overlapping and their centres are localised almost on the same spot. In particular, the cyan cluster is covered by the purple and green clusters.

An inspection of the documents in these four clusters showed that the cyan cluster was connected with the green one, while the orange cluster was linked with the purple one due to their textual interconnectivity. In the purple and orange clusters, the resources were more informative, discussing news, interviews and political facts

(cf. Figure 17a). The cyan and green clusters mostly consisted of explanatory texts and entries of online lexicons (cf. Figure 17b).
In total, the joined clusters would comprise 2,906 and 2,936 files, validating the k-means assumption that clusters contain roughly the equal number of observations.

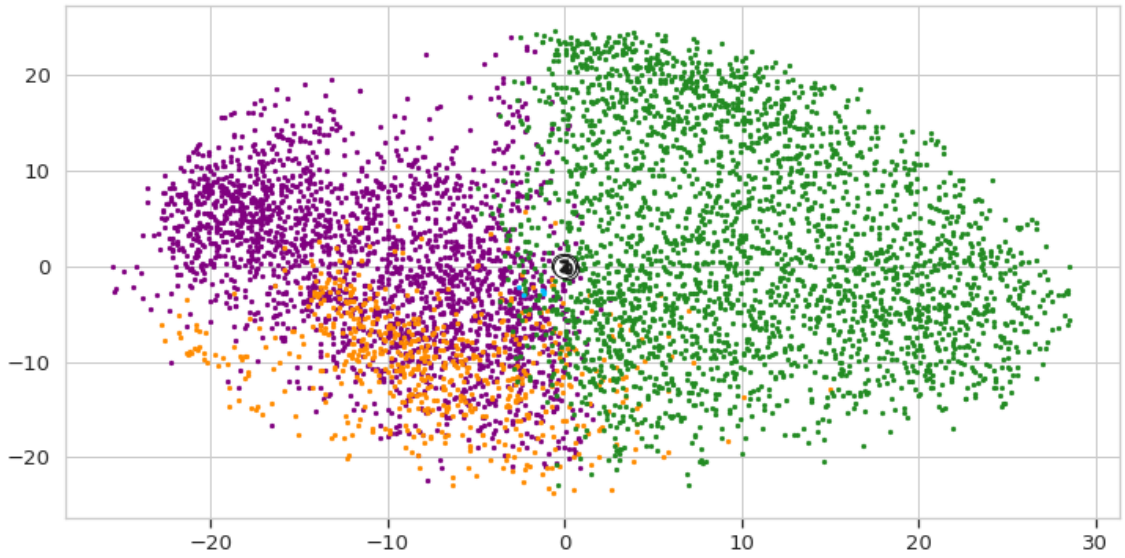


Figure 16: Scatter plot of the clusters of the complete dataset using Subset 2.

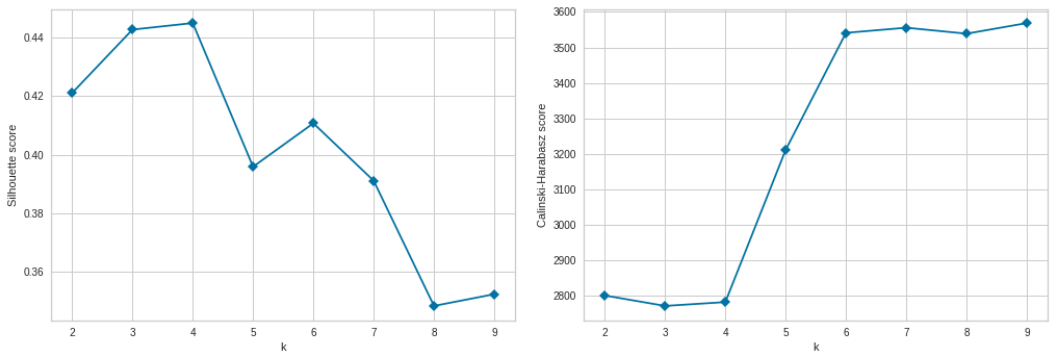
By comparison, the experiment using agglomerative hierarchical clustering suggested a three-cluster solution for the Subsets 2 and 4 (cf. Table 9, Figure 7). With regard to the second subset, Cluster 1 solely included entries of online lexicons. The resources in Cluster 2 consisted mainly of news articles and narrative texts. Cluster 3 was composed of explanatory texts and entries of online lexicons. When considering the genre of the texts, these outcomes suggest that Clusters 1 and 3 are deeply connected and may be analysed in combination, as for k-means. It could be, however, that these two clusters are defined by distinguishing features that account for a fundamental difference among textual complexity levels and, hence, that the cyan cluster in Figure 16 should not be merged with the green cluster.

To summarise, the k-means algorithm generally divided the texts into two clusters, as visible in Figures 14 and 18. The exploration of the files included in the clusters enabled me to assume that the textual genres have structural and typographic properties, which are so peculiar that they influence the clustering performance of the algorithm.

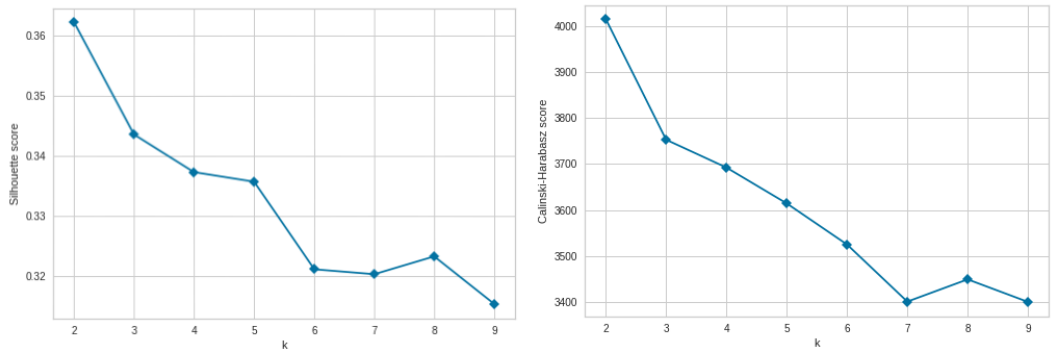
(a) Purple and orange clusters

(b) Green and cyan clusters

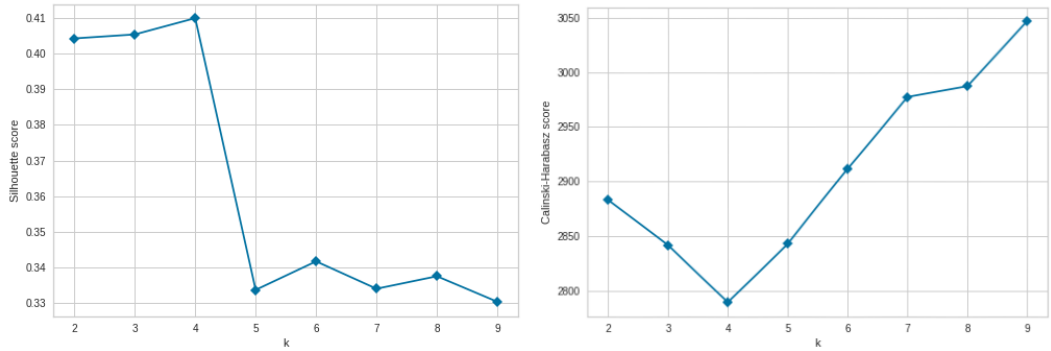
Figure 17: Example of a snippet of text for each cluster represented in Figure 16. Texts from `https://taz.de/`, `https://www.stadt-koeln.de/`,`https://hurraki.de/`, `https://www.bundesregierung.de/` (last accessed: 30 May 2019).



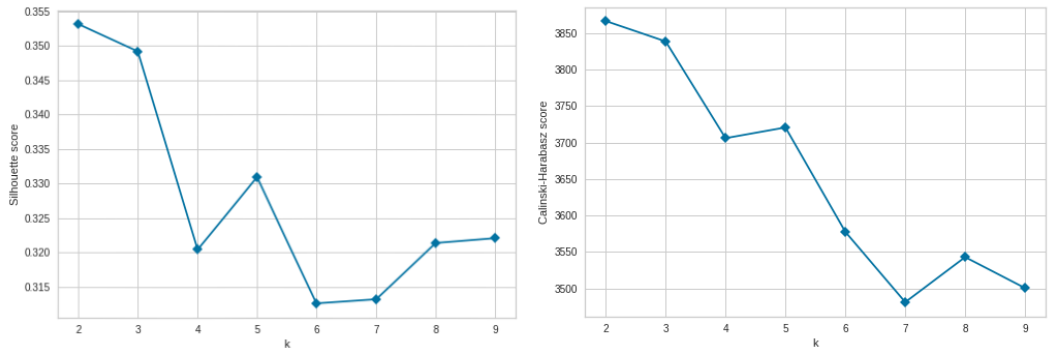
(a) Subset 2



(b) Subset 3



(c) Subset 4



(d) Subset 5

Figure 18: Plots of the silhouette coefficients and Calinski-Harabasz values for Subsets 2, 3, 4 and 5. Note that the highest value should coincide with the best partitioning. I did not display the plots of the intracluster SSE, because they did not differ from the plot in Figure 14b.

Webpages vs. PDFs

The results in Table 13 illustrate that the best partitioning of the webpages occurred when *k=2*, while the clustering algorithm in general separates the PDF files into three groups (*k=3*). However, the visualisation of the data showed that the highest silhouette coefficients and Calinski-Harabasz scores did not correspond to the best partitioning. For example, the first subset of the PDFs achieved a silhouette coefficient above 0.46 in line with *k=4* (cf. Table 13). Figure 19 shows that clusters 0 and 2, which consist of 26 and 29 data points, potentially include outliers. Their silhouette profiles have visibly different widths and lengths, suggesting a less reliable partitioning of the clusters. As observed in the analysis of the complete corpus, these types of clusters may contain resources with a special feature characterisation, indicating either too simple or too complex texts. The feature analysis of the four-cluster solution demonstrated that each cluster had distinctive properties. In particular, two clusters tended to be characterised by features denoting complexity and two clusters by features indicating simplicity. One cluster included texts with the highest rate of prepositions governed by genitive (0.0007 vs. 0.0001, 0.0003, 0.0004) and present-participle constructions (0.001), which were missing in the other three clusters. The second cluster was determined by the highest rate of negations (0.53 vs. 0.02, 0.27, 0.09) as well as by the lowest rates of foreign words (0.00004 vs. 0.0002, 0.00002, 0.0004), interjections (0.000005 vs. 0.00005, 0.0004 , 0), compound ratio (0.00004 vs. 0.0002, 0.0004, 0.0001) and postpositions (0.000008 vs. 0.0002, 0.00005, 0.0001). In the third cluster, interjections, semicolons, indirect speech constructions, genitive objects and words difficult to perceive were missing, while the variables for paragraphs and images obtained on average a slightly higher rate than in the other three clusters (0.05 and 0.03 vs. 0.02, 0.006; 0.03, 0.01; 0.04, 0.02). Finally, the fourth cluster was determined by a high rate of special characters (0.19 vs. 0.00001, 0.007, 0.0005) and words in bold (0.78 vs. 0.04, 0.2, 0.09) and by the absence of subjunctives and present-participle constructions.

An examination of the documents in these clusters (cf. Figure 19) showed that the blue cluster was connected with the green one, while the yellow cluster was linked with the black one. The resources in the blue and green clusters were narrative texts. The yellow and black clusters consisted of texts ranging from instructions, guidelines, lists of job descriptions to portrays of the political parties. As already stated in the experiment with the complete dataset, I assume that the properties characterising each textual genre are so peculiar that they can affect the clustering performance of the algorithm.

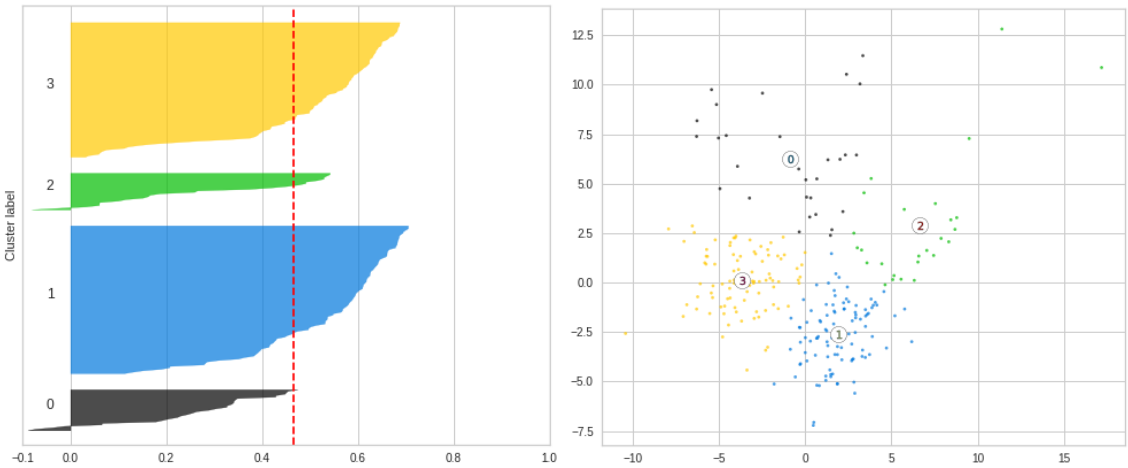
Overall, a size of k equal to 2 or 3 was a good cluster solution for the text subsets using the k-means algorithm. This result was similar to the solution suggested by

	Subset 1		Subset 2		Subset 3		Subset 4		Subset 5	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
Webpages										
2	**0.36**	**3747.3**	**0.542**	2585.8	**0.369**	**3740.1**	0.40	2722.8	**0.348**	**3609.7**
3	0.339	3507.4	0.401	3203.8	0.34	3505.8	**0.402**	**2723.0**	0.34	3592.8
4	0.338	3447.1	0.41	**3304.2**	0.339	3452.2	0.345	2652.7	0.318	3480.47
PDFs										
2	0.424	169.1	**0.63**	142.6	0.424	178.7	0.553	142.6	0.44	214.8
3	0.46	**202.6**	0.456	179.6	**0.456**	**218.9**	**0.561**	**233.3**	**0.453**	**245.9**
4	**0.466**	201.8	0.479	**186.2**	0.409	208.4	0.396	224.5	0.383	223.0

Table 13: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) using k-means on webpages and PDF files.

the analysis of the complete dataset. Inspecting the features defining the clusters demonstrated that linguistic differences were evident among groupings of texts.

In general, it is important to note that k-means is highly sensitive to outliers and the results can be severely skewed if the dataset includes texts scraped from the web, without quality assurance, as the resources in my corpus.



(a) Silhouette profile

(b) Scatter plot

Figure 19: Visualisation of the silhouette profile and scatter plot of the best k for the first feature subset of PDF files.

3.5.2.2 With Feature Agglomeration

The aim of this subexperiment was to investigate whether feature agglomeration of sparse features could improve the clustering performance. As for the subexperiment in Section 3.4.2.2, I applied the algorithm with different sizes of `n_clusters`; the algorithm trained with the hyperparameter *n_clusters=3* performed better than the algorithm trained with *n_clusters=5*.

Complete Dataset

By comparing the line plots in Figures 18 and 20, reducing the feature space through feature agglomeration did not lead to an improvement in the silhouette coefficients. On one hand, the silhouettes tended to delineate a relatively smooth curve in the Subsets 1, 2 and 5, yielding the best partitioning in line with *k=2*, which corresponded to the highest score. On the other hand, the values for the Subsets 2 and 4 were lower than the same values in the analysis without feature space reduction. Subset 2 performed better without feature agglomeration also in the hierarchical clustering analysis (cf. Section 3.4.2.2).

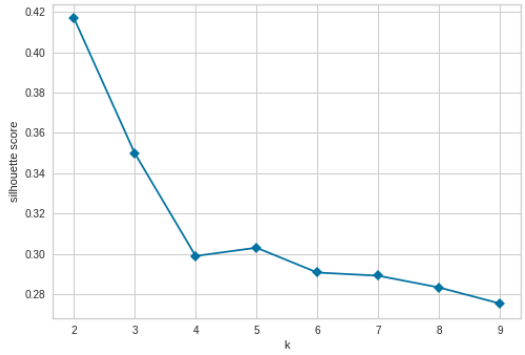
These results suggested that the feature agglomeration process was not advantageous to the performance of the k-means algorithm. Future research may investigate the causes of these findings.

Webpages vs. PDFs

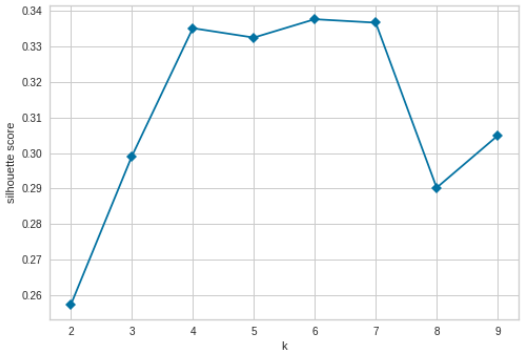
Table 14 summarises the results of k-means clustering on webpages and PDFs, after grouping the features into three clusters. Compared to Table 13, the feature agglomeration step improved the partitioning for the websites by making the gap in the values among different number of clusters more evident. This suggested that the difference of the silhouette scores between *k=2* and *k=3* was greater than in the analysis without feature agglomeration (0.36 - 0.339 vs. 0.399 - 0.345). It is worth noting the improvement of 23% in the Calinski-Harabasz indices among all five feature subsets for the webpages. However, for PDFs, the evaluation metrics achieved diverse values in the Subsets 1 and 2.

On the whole, k-means obtained a more unclear partitioning of the texts than the hierarchical clustering algorithm (cf. Table 12). The feature agglomeration process did not improve the partitioning.

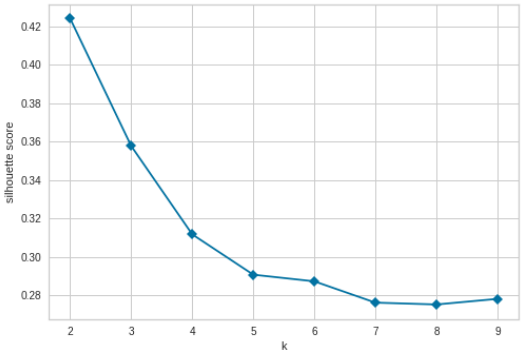
To sum up, as previously reported, k-means have some drawbacks [Raval and Jani, 2016; Sonagara and Badheka, 2014]. The algorithm assumes that the clusters have a spherical shape and that each of them contains roughly the equal number of observations.



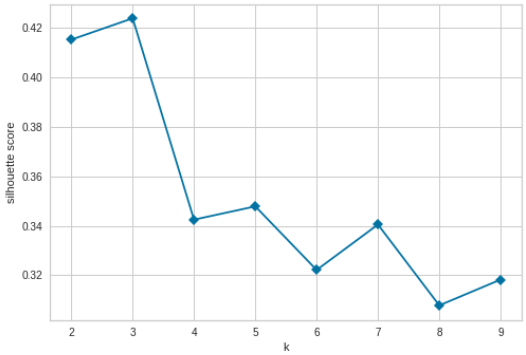
(a) Subset 1



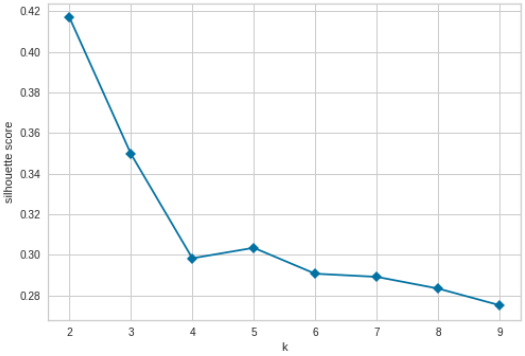
(b) Subset 2



(c) Subset 3



(d) Subset 4



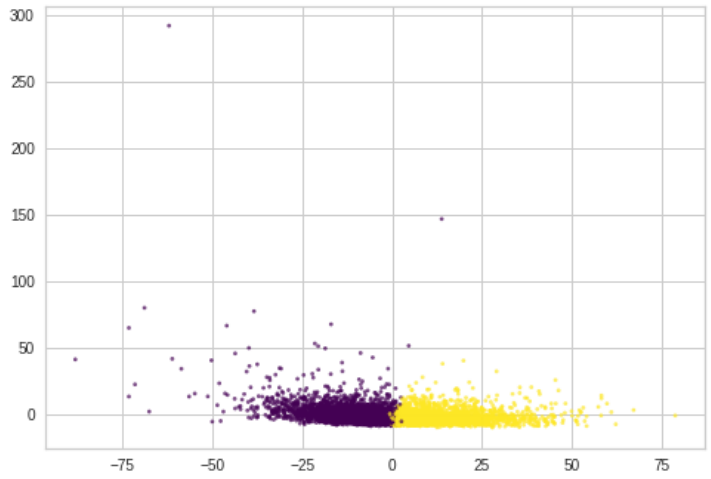
(e) Subset 5

Figure 20: Plots of the elbow method for different number of clusters using the silhouette coefficients after feature agglomeration.

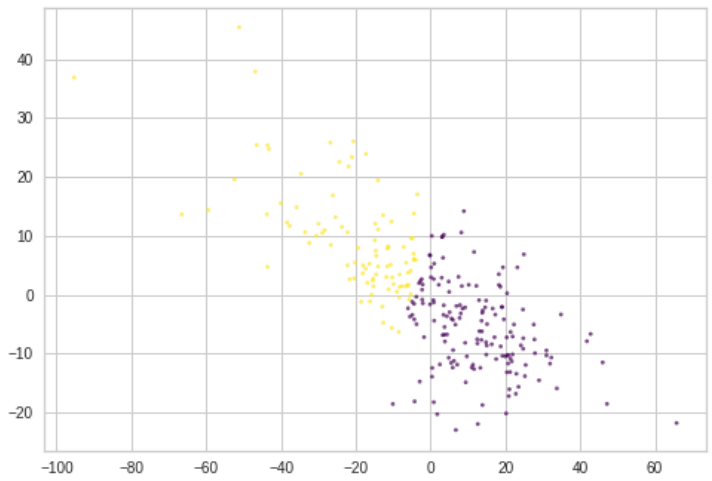
Figure 21 shows the scatter plots of the webpages and PDFs using k-means after agglomerating the complete feature set. Having set the hyperparameter `n_clusters` to 2, the algorithm identified two clusters. It is difficult, however, to distinguish between two different groups of data points. Together with Figures 16 and 19, these plots demonstrate that the assumptions of k-means are violated in my dataset. For these reasons, I can argue that the k-means algorithm is not suitable for clustering my corpus in simplified German.

	Subset 1		Subset 2		Subset 3		Subset 4		Subset 5	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
					Webpages					
2	**0.399**	**4466.6**	**0.531**	**5122.5**	**0.41**	**5098.7**	0.413	**3510.4**	**0.406**	**5056.3**
3	0.345	3884.0	0.473	5084.9	0.344	4547.9	**0.423**	2878.6	0.337	4469.5
4	0.31	3347.5	0.459	4971.0	0.305	4050.9	0.345	2892.9	0.326	3961.6
					PDFs					
2	0.412	**210.2**	**0.578**	183.3	0.407	218.6	0.418	137.9	**0.477**	**281.1**
3	0.387	201.9	0.446	191.0	**0.420**	**227.5**	**0.461**	**160.1**	0.463	226.6
4	**0.424**	198.6	0.465	**206.8**	0.352	200.8	0.348	149.9	0.398	224.3

Table 14: Comparison of the silhouette scores (Sil) and Calinski-Harabasz (CH) indices using k-means on webpages and PDF files after feature agglomeration.



(a) Webpages



(b) PDFs

Figure 21: Scatter plots using k-means after agglomerating Subset 1.

3.6 Experiment 3: Clustering TopEasy News Text Collection

In the previous experiments, the noisy nature of the corpus limited the performance of the algorithms. The term *noisy* refers to the quality of the texts, which were entirely crawled from the web and potentially contain various errors, such as spelling and segmentation errors.
In this experiment, I repeated the two proposed approaches on a further text collection in simplified German (cf. Section 3.1.2). Since the texts in this collection were written by a unique professional translation provider, I assumed a consistent high-quality of the material. The texts were labelled at two different levels of simplified German, but I did not consider this labelling a validation method. However, if the labels A2 and B1 were correctly assigned, my clustering models would group the texts into two clusters.

3.6.1 Method

I conducted two experiments on the "TopEasy News" text collection. For each text, all features were extracted and stored in a file as comma-separated values. Prior to PCA, the feature values were normalised and scaled. In the first experiment, I repeated the agglomerative hierarchical clustering method applied in Section 3.4.1. The second experiment was based on the k-means method used in Section 3.5.1. In both experiments, I compared the performance under five different feature subsets. In this section, I summarise the experimental settings.

Agglomerative Hierarchical Clustering

I made use of the `linkage` function from the `Scipy` library. Instead of running the algorithm on different metric-linkage combination, I directly chose the pair cosine-average, which yielded a CCC of 0.738. The optimal size of k was selected according to the following methods: (1) by inspecting the dendrograms; (2) by recursively running the algorithm and (3) calculating the silhouette score and Calinski-Harabasz index for each iteration. The analysis was repeated among the five feature subsets and after the feature agglomeration step. Note that I did not apply PCA prior to feature agglomeration, since both methods aimed at reducing the feature space.

K-means

Differently from hierarchical clustering algorithm, I made use of the `scikit-learn` library to create k-means models. The centroids were initialised by setting the pa-

rameter `init=k-means++` [Arthur and Vassilvitskii, 2007]. To determine the optimal k, I recursively ran the algorithm with the parameter `n_clusters` in a range from 2 to 9 and calculated the silhouette score and Calinski-Harabasz index for each iteration. As additional evaluation method, I visualised the silhouette profiles and the scatter plots. I repeated this approach among the five feature subsets and after the feature agglomeration technique.

3.6.2 Results

3.6.2.1 Without Feature Agglomeration

Agglomerative Hierarchical Clustering

Table 15 shows the results of this experiment, with the best entries highlighted in bold. Interestingly, the results are not consistent among and within the feature subsets. The clustering analysis achieved the best silhouette and Calinski-Harabasz scores with a k between 2 and 5. In the second and fifth subset, the algorithm suggested that k=*4* was the optimal number of clusters. The values for the third and fourth subsets implied an optimal clustering between k=*3* and k=*4*. These results are corroborated by the visualisation of the hierarchical trees in Figure 34 (Appendix D), which compares the dendrograms of all five feature subsets. Each dendrogram shows a tripartite structure, in which every main cluster includes one to two more fine-grained subclusters.

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
2	0.507	417.9	0.666	721.6	0.607	592.6	0.552	405.7	0.567	503.8
3	0.637	**591.3**	0.706	871.0	0.637	**569.3**	0.580	**558.6**	0.621	575.8
4	0.633	583.2	**0.81**	**1147.2**	**0.650**	547.4	**0.624**	556.3	**0.678**	**602.6**
5	**0.652**	586.3	0.77	986.6	0.648	523.4	0.591	494.62	0.643	514.0

Table 15: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) among the five feature subsets.

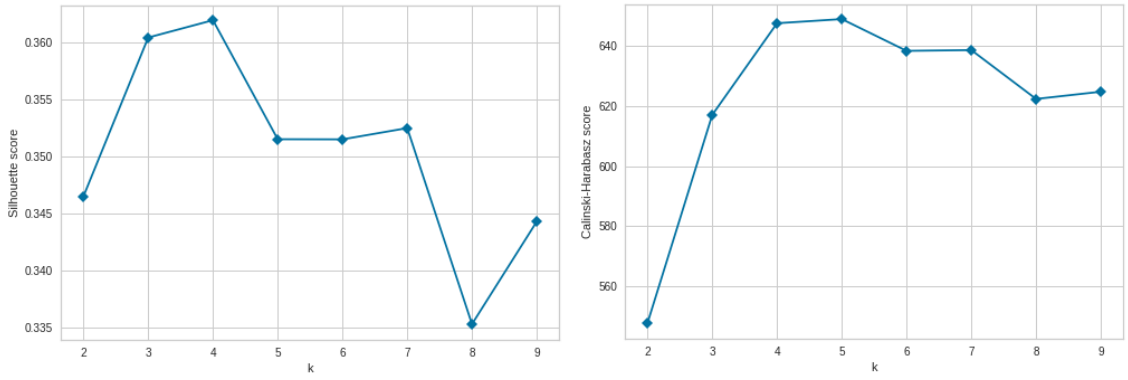
When considering the features characterising the clusters, the mean ratios are so low (e.g., 0.000005 for pronouns) that any evident differences among clusters are not observed. A more detailed statistical analysis is necessary to determine which linguistic features can bring the algorithm to cluster texts together. Only the outcome of the Subset 2 is defined by a clear distinction in the feature distribution,

which I discuss in Chapter 4 in relation to research question 3 (cf. Section 4.2 and Figure 30).

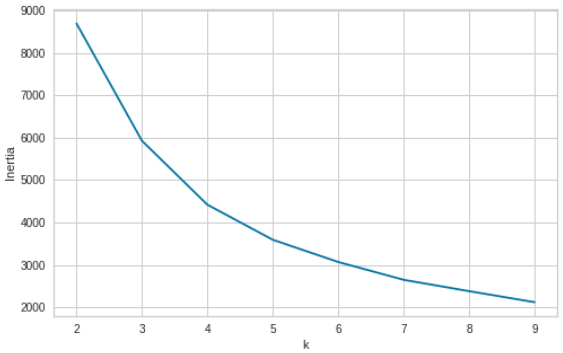
In summary, while the evaluation metrics and the visualisation of the dendrograms demonstrate a clustering of texts, the difference in the mean ratio of the features is too small to achieve a complete overview of the clustering. It is worth noting that in this analysis the textual genre cannot affect the partitioning, since the whole text collection consists of news articles.

K-means

Figure 22 illustrates the silhouette coefficients, the Calinski-Harabasz and the intracluster SSE scores for the first feature subset. The silhouette and Calinski-Harabasz metrics reached their highest value in line with k=*4*, while the intracluster SSE displayed a smooth line without a distinct "elbow", meaning that the magnitude of the value decrease did not drop. The first subset achieved diverse results also in the previous experiment with agglomerative hierarchical clustering.



(a) Silhouette score and Calinski-Harabasz score

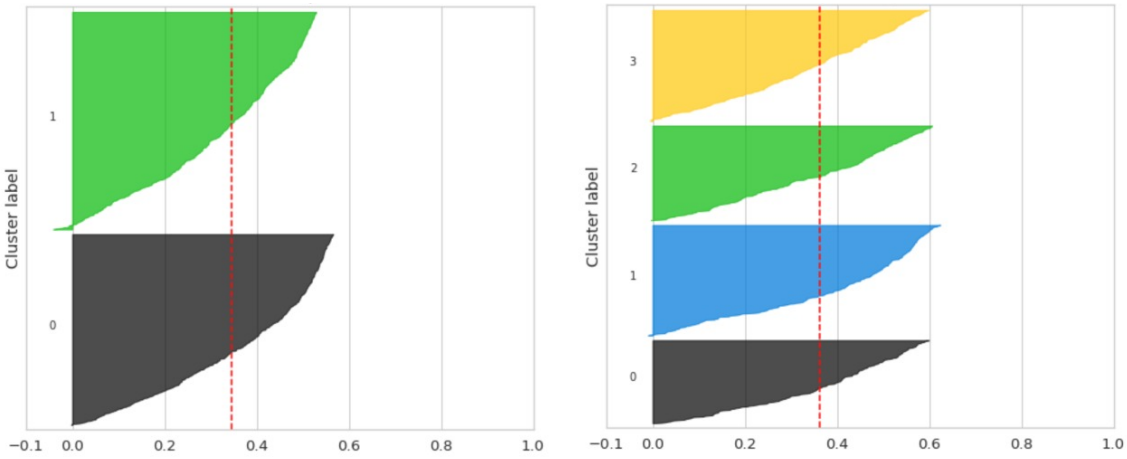


(b) Inertia

Figure 22: Plots of the elbow method for different number of clusters using the feature Subset 1.

The visual inspection of the silhouette profiles for k equal to 2 or 4 shows that both sizes may be a good partitioning of the dataset, since the profile forms tend

to be similar in height and width among clusters (Figure 22). The reason for this partitioning can be attributed to the sparse and heterogeneous linguistic information included in the feature set. The combination of all features potentially led the clustering model not to converge on a specific size of k, indicating that k-means did not suit the dataset or the feature set was not representative of the resources.



(a) k=2 (b) k=4

Figure 23: Plots of the silhouette scores for two and four clusters for Subset 1.

The well-shaped silhouette for k=2 in the figure above could lead one to think that the two clusters coincide with the labelling A2 and B1. The scatter plot in Figure 24 visibly shows that the files were distributed in the two clusters independently of their language level label.

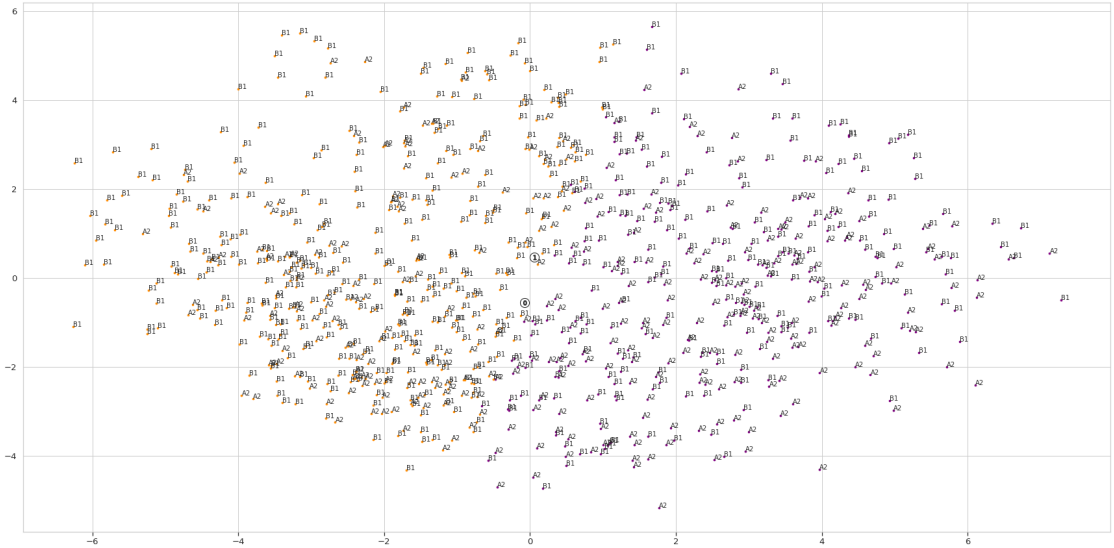


Figure 24: Scatter plot of files considering the feature Subset 1.

With regard to the other feature subsets, the evaluation metrics achieved highly varying values (cf. Table 16) suggesting different partitioning of the dataset in comparison to Table 15. Only the scores of the second subset corroborated the results obtained in the previous experiment, namely that 4 is the optimal size for k considering the surface features.

On the whole, for all five feature subsets, the best partitioning of the dataset occurred when the size of k was between 2 and 4. As shown in Figure 24, the two-cluster solution does not correspond to the language level labels A2 and B1. The comparison between the *capito* features defining these two language levels and the features characterising the resulting clusters could be beneficial to detect any crucial similarity or dissimilarity to consider for further analyses of simplified German.

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
2	0.346	547.8	0.444	722.4	0.359	609.1	0.323	455.2	**0.376**	**672.8**
3	**0.361**	617.0	0.508	998.5	**0.365**	610.8	**0.38**	**644.1**	0.369	665.5
4	0.36	648.1	**0.551**	**1238.3**	0.348	604.7	0.336	594.6	0.340	653.9
5	0.351	**648.8**	0.49	1144.1	0.345	**618.6**	0.336	592.4	0.335	632.2

Table 16: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) among the five feature subsets using k-means.

3.6.2.2 With Feature Agglomeration

The purpose of this analysis was to examine agglomerative hierarchical clustering and k-means algorithms after feature agglomeration. Table 17 shows the results of this subexperiment trained with `n_clusters=3`. For both clustering approaches, silhouette coefficients and Calinski-Harabasz scores on average increased by almost 16% and 62.6%, respectively (cf. Tables 15 and 16). From the results, I obtained the following observations: (1) The feature agglomeration step visibly enhanced the values of the evaluation metrics, especially the Calinski-Harabasz index; (2) After feature agglomeration, both algorithms tended to group the texts into two clusters. These observations were evident also in the subexperiments run on the noisy corpus of simplified German (Sections 3.4.2.2 and 3.5.2.2).

Overall, the feature agglomeration process helped to improve the partitioning of the texts, according to the scores obtained by the evaluation metrics. By examining the agglomerated features I found that several features were missing in the texts of this

collection. In the discussion, I come back to this issue addressing research question 3 (cf. Section 4.3).

	Subset 1		**Subset 2**		**Subset 3**		**Subset 4**		**Subset 5**	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
Agglomerative hierarchical clustering										
2	**0.748**	**1149.2**	0.698	**1106.4**	**0.675**	**797.6**	**0.595**	**596.9**	**0.646**	**715.7**
3	0.687	686.1	**0.705**	801.0	0.645	585.8	0.575	513.7	0.566	488.6
4	0.739	569.5	0.684	537.0	0.657	511.2	0.562	455.1	0.505	398.8
K-means										
2	**0.491**	1244.9	**0.474**	**1156.6**	**0.427**	**901.5**	**0.358**	**631.3**	**0.404**	**773.4**
3	0.426	1244.5	0.434	1119.7	0.343	773.6	0.316	577.3	0.319	628.2
4	0.404	**1262.4**	0.376	1007.2	0.313	695.2	0.321	549.9	0.308	583.7

Table 17: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) among the five feature subsets using agglomerative hierarchical clustering and k-means after feature agglomeration.

3.7 Summary

This chapter has provided a description of the clustering analysis I performed to investigate evidence of multiple complexity levels in simplified German. Firstly, I have presented the resources in simplified German, which are annotated not only at a linguistic level but also with structural and typographic information. The first and main resource for my analysis is a new corpus in German/simplified German, which consists of texts crawled from the web. The second dataset is a collection of news articles in simplified German, written by a unique translator provider.

The experiments consisted of exploring empirical grounding for multiple complexity levels among several feature combinations and subsets (cf. Sections 3.4 and 3.5) and variable data quality (cf. Section 3.6). This implied the analysis of the clustering results and the investigation of the features defining those results. In the experimental approaches, I used two clustering methods (agglomerative hierarchical clustering and k-means) and two dimensionality reduction techniques (PCA and feature agglomeration).

In general, the results showed that the agglomerative hierarchical clustering algorithm performed better than the k-means algorithm on both datasets. Agglomerative hierarchical clustering achieved more consistent results than k-means in relation

to the complete dataset, subsets of the same and among different feature combinations. The dendrograms visibly indicated a clustered structure, while the k-means algorithm did not divide the datasets into distinctive clusters as shown in the scatter plots. This can be explained by an interconnectivity in their linguistic and structural characteristics, which the second algorithm cannot represent in the best way, as described in the context of Figures 16 and 19.

Given the noisy and hence challenging nature of the first dataset, my experimental approaches achieved promising performance, especially when considering that I exploited not only linguistic but also structural and typographic characteristics, individually and separately. The subsequent chapter discusses the results in the context of the research questions posed in Section 1.2.

4 Discussion

So far, I have performed and evaluated separately three clustering experiments and several subexperiments with texts in simplified German. I have demonstrated that texts in simplified German can be divided into various clusters described by dissimilar mean ratios of the features, but I have not looked for an explanation accounting for this clustering. In the following sections, I explore and discuss the results, so as to give answers to the research questions outlined in the introduction (cf. Section 1.2).

4.1 Research Question 1

Is there empirical evidence of multiple complexity levels in existing simplified German texts?

On the whole, the analytic results show that there are different characteristics among texts that lead to their clustering in separate groups. This trend is more remarkable with a more controlled input, such as "TopEasy News" text collection, and after the reduction of the feature space. As explained in the introduction (cf. Section 1), it is worth noting that multiple complexity levels in simplified German do not imply a correspondence with CEFR linguistic levels. The different complexity levels may coincide with the needs of the target users. Therefore, further investigations have to be performed to determine if the obtained clusters actually correspond to multiple complexity levels in simplified German. This involves a future significance testing for clustering, even if it is widely considered an arduous and problematic task [Evert et al., 2019].

The box plots of the features gave me a detailed insight into the feature distribution in order to make further progress with the research question. Figure 25 visualised the box plots of six of the surface features (Subset 2) based on the clustering suggested by the agglomerative hierarchical approach (cf. Table 9). The difference between clusters is indicated by their higher or lower position on the plane. For example, the box plots for the variable *dot* suggest that the second cluster contains texts

with a lower count of full stops than the first and third clusters. The box plots for the feature *comma* indicate that its rate increased across the three clusters. As introduced in the results in Section 3.4.2.2, these findings verify the importance of distinguishing among punctuation marks, in contrast to the approach adopted by Pilan and Volodina [2018] in their ARA system (cf. Section 2.3.1). Interestingly, the first cluster consists of texts that follow the rule *one-sentence-per-line*, have a low frequency of the variable *comma* and a high number of paragraphs. These characteristics are crucial properties of simplified texts. Texts in the first cluster do not include images or charts, even though several outliers exist. According to a study performed by Bock [2018], target readers tend to disregard images or pictures if included in large quantity in a text. Yet they tend to focus on single explicative images (photographs better than pictograms) underlining the relevance of a visual element in a text.

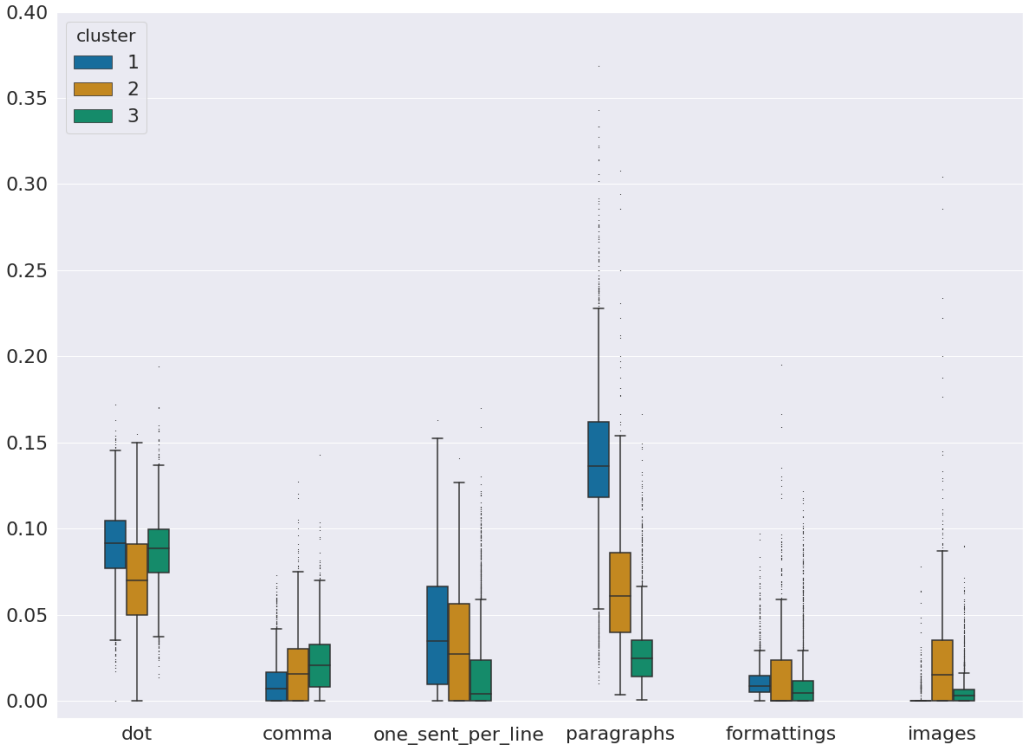


Figure 25: Box plots of six features of the surface set (Subset 2). Labels for: frequency of full stops, commas, "one-sentence-per-line", count of paragraphs, frequency of typographic information, count of images and charts.

Figure 26 compares three clusters among six features from the lexical and semantic group (Subset 4), which were based on word lists (Section 3.2.1), with the exception for the bi-logarithmic TTR [Vajjala and Meurers, 2012]. In isolation, the frequency of words may not be indicative of the language level, because variables, such as

education level and experience of the target readers, influence the understanding of words [Bock, 2018]. In general, the frequency of words still provides a reference for a specific language level. This is confirmed by the relative distribution of the variables throughout the clusters in the Figure 26. The data highlights a correlation between the distribution of the most frequent words in German (*mostFreqDereko*) and simple specialised words (*a1_sachbereich*), which are easily perceived and understood. Focusing on the category of specialised words, a decreasing trend in their usage among clusters and levels of perception can be seen. Specialised words at a level B1 are not as widely employed as simple specialised words, namely at a level A1. The feature *lastFreqDereko* and the last three features on the figure (*a2_sachbereich*, *b1_sachbereich*, *BilogTTR*) display comparatively short box plots suggesting that overall texts in each cluster have a high level of agreement on that specific variable. In particular, the low level of the variable *lastFreqDereko* reveals that the least frequent words in the German reference corpus DeReKo [Institut für Deutsche Sprache, 2014], generally perceived as difficult, are almost completely absent from all the texts in the corpus.

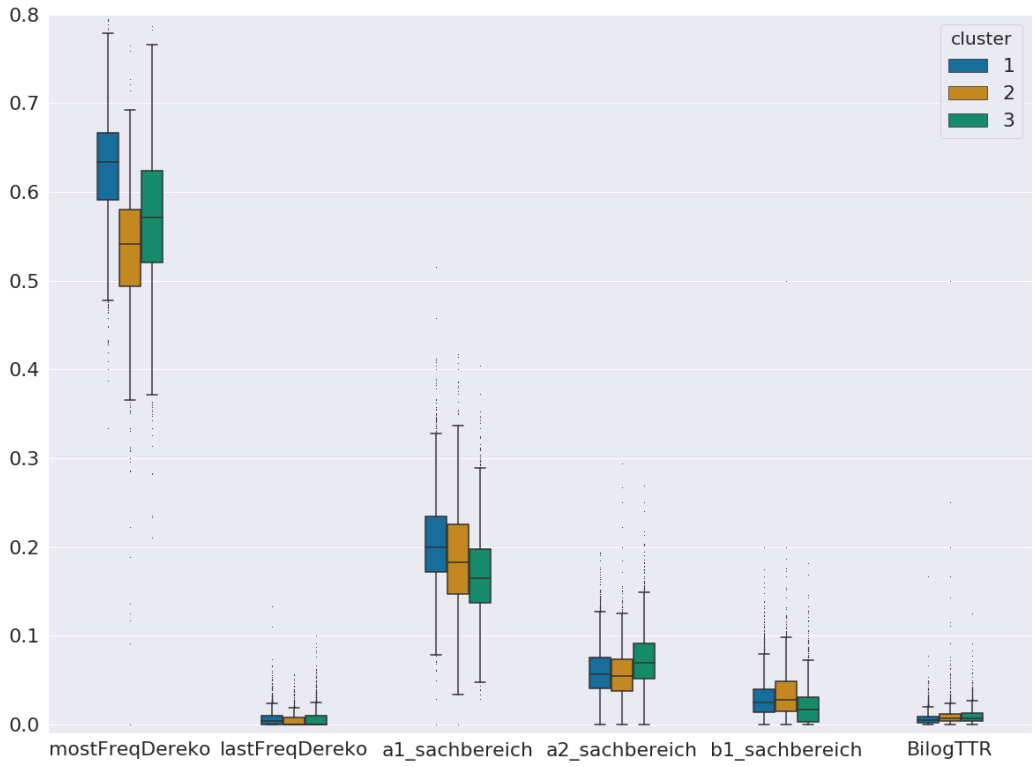


Figure 26: Box plots of lexical and semantic features (Subset 4). Labels for: count of the most frequent words, count of the least frequent words, count of specialised words at different perceptive levels (level A1, A2 and B1), bi-logarithmic type-token ratio.

Six features from the morphological, morpho-syntactic and syntactic feature group (Subset 5) are displayed in Figure 27. In the three clusters, the number of nouns (*NN*) is inversely proportional to that of verbs (*V*). Texts in Cluster 1 use more verbs than nouns, suggesting that their linguistic structure is simpler than the texts in the second or third cluster. Cluster 2 includes texts focusing on objects or concepts, since verbs (events, actions, etc.) have been turned into nouns (concepts, things, etc.) following the linguistic process of nominalisation. Interestingly, Cluster 1 includes a large number of pronouns, negation and subordinate clauses. A deeper analysis of these features is necessary to define their categories. As discussed in Chapter 2, there are pronouns as well as subordinate clauses which can simplify or complicate text readability and comprehension.

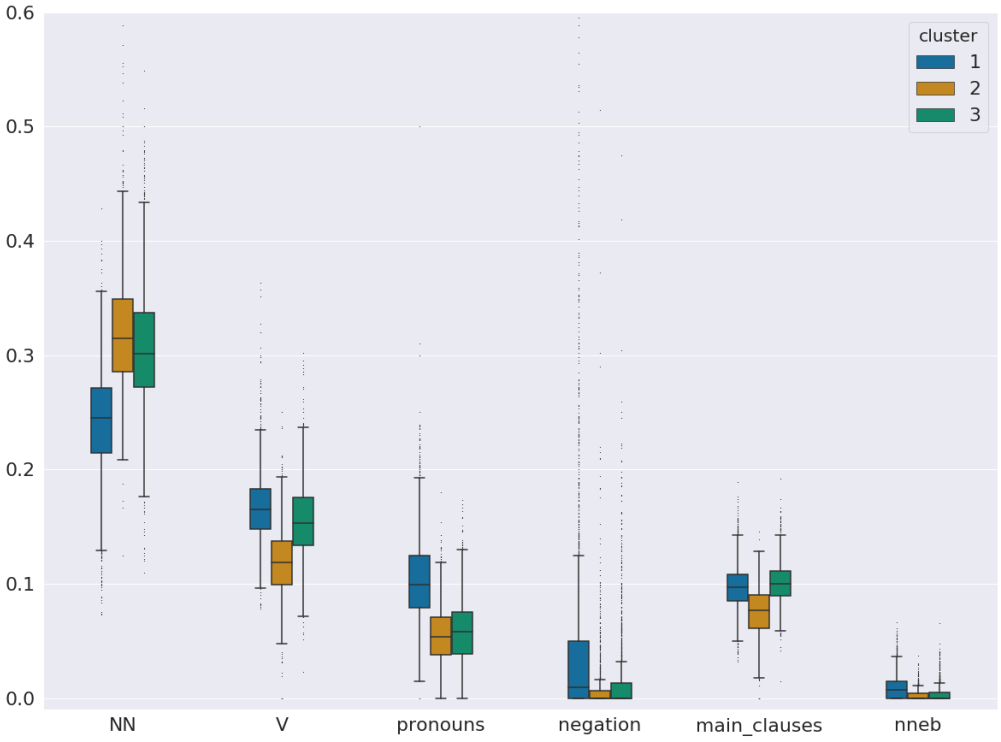


Figure 27: Box plots of morphological, morpho-syntactic and syntactic features (Subset 5). Labels for: frequency of nouns, verbs, pronouns, negation, count of main clauses, count of subordinate clauses.

For the configuration "webpages vs. PDFs", eleven features from the deeper feature set (Subset 3) are compared in Figure 28. I created these box plots after the feature agglomeration step, which performed the clustering of texts into two groups for both subsets (cf. Section 3.4.2.2). What stands out in this figure is the high variability of the feature *negation*. Despite the large number of outliers, webpages (top) include a small number of words indicating negation. The negation rate in the PDF files (bottom) is visibly larger and extremely variable in the clusters. This outcome relates

to the claim that negation expressed through *nicht* and *kein* (en: *no, not*) is easy to understand [Bock, 2018]. Future research can consider double negation and negative constructions, such as *weder...noch* (en: *neither...nor*) and negative words in order to capture all the complexity shades of this variable. It is important to note that the algorithm, which does not remain consistent across different runs, randomly assigned the numbering and the colouring of the clusters.

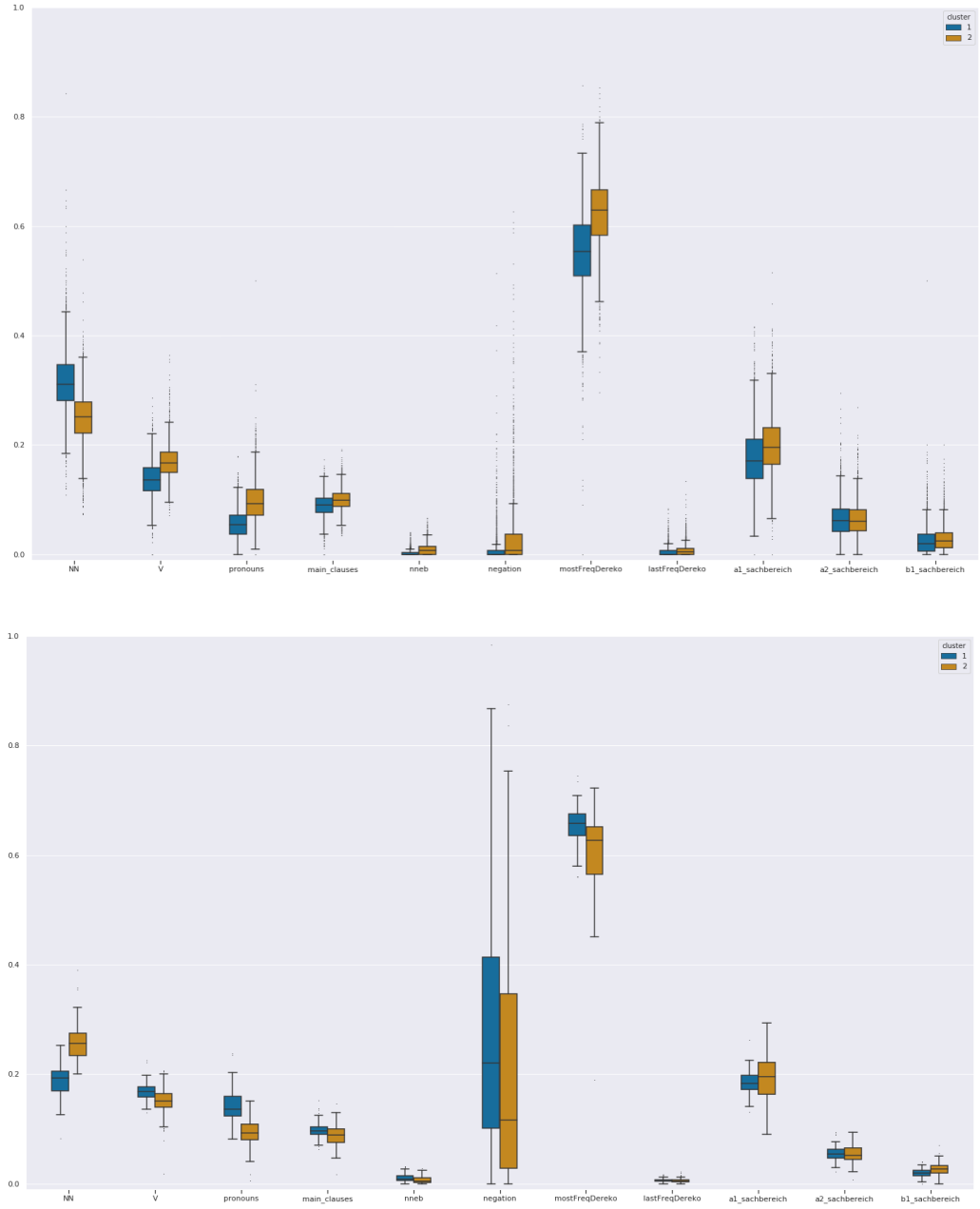


Figure 28: Box plots of deeper features (Subset 3): webpages vs. PDFs. Labels for: frequency of nouns, verbs, pronouns, main and secondary clauses, negation, count of the most and the least frequent words, count of specialised words at different perceptive levels (A1, A2 and B1).

4.2 Research Question 2

Can an unsupervised machine learning approach, such as cluster analysis, provide empirical evidence of multiple complexity levels?

Before selecting the two investigated approaches, I tested several unsupervised learning algorithms. Amongst others, I performed experiments with density-based spatial clustering of applications with noise[1] (DBSCAN), mean shift[2] and Gaussian mixture models[3] (GMM).[4] The final choice of agglomerative hierarchical clustering and k-means algorithms was made as a consequence of their wide usage in document clustering research [Aggarwal and Zhai, 2012]. Yet they might not be the most appropriate algorithms for my dataset. On one hand, it is worth noting that k-means clustering is conceptually not suitable for this research analysis, because it requires a predetermination of the number of clusters, which contradicts the aim of a fully inductive procedure that should be at the basis of this investigation. On the other hand, the varying results also showed that agglomerative hierarchical clustering and k-means were not always suitable models to cluster the datasets for each feature subset. In Table 9 and Figure 18, the silhouette coefficients display for each subset a second relevant value, which is almost as good as the chosen one in line with *k=3*.

At this point, the appropriateness of the evaluation metrics can also be discussed. In my analysis, the range of possible evaluation approaches was limited due to unlabelled data. This means that there was no ground truth with which to compare the results. The relative success of all experimental models was hence evaluated against internal evaluation metrics, which used information on computed clusters to assess their well-separation, compactness or density [Halkidi et al., 2001]. As presented in the results of each experiment (cf. Sections 3.4.2, 3.5.2 and 3.6.2), there was not a clear k value on which the evaluation metrics agreed unequivocally, even if the final decisions were similar. Liu et al. [2010] explain the limitations of internal evaluation metrics in different application scenarios. In particular, the silhouette score is influenced by the presence of subclusters (e.g., Subset 1 in Table 9), while the Calinski-Harabasz index is significantly affected by skewed and noisy data (e.g., Table 10). For this reason, in my analysis, the visualisation of dendrograms and plots played a significant role while determining the optimal number of clusters and

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html` (last accessed: 15 May 2019)

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html` (last accessed: 15 May 2019)

[3] `https://scikit-learn.org/stable/modules/mixture.html` (last accessed: 15 May 2019)

[4] I did not examine these approaches in depth. Therefore, I did not include the results in this thesis.

interpreting the results.

Another drawback of unsupervised algorithms ought to be considered, namely their sensitivity to outliers. Outliers can range from being irrelevant to being explicative of the data. In my experimental approaches, I intentionally did not detect the outliers, so as not to underestimate the variance or reduce the small dataset. However, I cannot exclude the presence of outliers since my dataset was entirely crawled from the web. As described in Section 3.5.2.1, retrieving and examining the files in these clusters would help to determine their nature. For instance, file `56.tcf` in my corpus was separated from all clusters during several experiments. An inspection of the content enabled me to verify its unusual construction: The text consists of a list of contact persons, including address, email address, phone and fax numbers. Outliers can hence be decreased by carefully controlling the corpus and deleting texts that do not cover a certain linguistic content during the preprocessing of the dataset. Otherwise, they can be detected during the clustering step. In agglomerative hierarchical clustering, outliers can be found by observing the dendrogram, since they tend to connect to one cluster lastly. In k-means, their identification is not so straightforward. A simple technique relies on the visualisation of the silhouette profiles and the scatter plot of the whole dataset after clustering (cf. Figure 19).

All this considered, after comparing the visualisation of dendrograms and scatter plots, I observed that agglomerative hierarchical algorithm led to more consistent and well-separated clusters than k-means. In addition, the feature analysis in Section 4.1 proves that the obtained clusters correspond to agglomerated texts at different complexity levels. These findings confirm that unsupervised learning, such as cluster analysis, can provide evidence of multiple complexity levels. Future work might deeply investigate the above-mentioned clustering algorithms, or new emerging algorithms, and focus on the detection of outliers.

4.3 Research Question 3

Are linguistic features suitable for this kind of analysis?

As mentioned in Section 3.2, clustering approaches highly depend on the feature set designed for the input of the analysis. In this thesis, I used simple linguistic features, because the clustering results have to be interpreted to create a framework of complexity levels, in which precise linguistic characteristics describe each level. Cha et al. [2017] proved that more sophisticated features, such as word and paragraph embeddings, could improve clustering results in readability assessment. However,

the inclusion of such features may complicate the interpretation step.

The feature set presented in Section 3.2 intends to be representative of simplified German from a linguistic and structural point of view. A correlation matrix of each feature subset (e.g., Figure 29) enabled me to discover variables that I discussed in the related literature but that were extremely sparse or even absent in the dataset. The variable *passive voice with agent* is one of these. Words in *italics* do not appear in the texts of the webpages, while the PDFs do not to contain any *semicolon*. Interestingly, the *boldface* type and *negation* are present among PDFs in large quantity. These two features achieved a correlation of 0.142, while the feature of *negation* negatively correlated with all other variables. This is because words indicating negation are usually marked in bold in simplified language, as Example 4.1 shows[5]:

(4.1) Die Arbeit darf **nicht** schädlich für die Gesundheit sein.
*The work must **not** be harmful to health.*
Es ist **kein** Platz frei.
*There is **no** free place.*

In the "TopEasy News" text collection, the following features are missing: *exclamation* and *question marks*, *semicolons*, *images*, *typographic information* (bold and italics), *answering particles*, *interjections*, *present-participle constructions* and some types of subordinate clauses. The textual genre, namely news articles, and the small number of texts limited the types of features. It is unlikely that news articles contain features such as questions, answering particles and interjections. I observed the impact of the textual genre on the feature and file distribution also in Sections 3.4.2.1 and 3.5.2.1.

By considering the complete dataset, this research question should be reformulated into a more specific one: "Are linguistic, structural and typographic features suitable for this kind of analysis?". As explained in Section 3.1.1, Battisti and Ebling [submitted] hypothesise typographic and structural aspects of texts to be highly predictive of simple vs. complex texts. With regard to the "TopEasy News" text collection, an examination of the clustering based on surface features (Subset 2) showed that the red cluster in Figure 34b only included texts at a level A2.[6] Among the features, *one-sentence-per-line* was the most salient variable for this respective cluster. This aspect can be attributed to two distinct reasons: (1) Webpages have been constructed differently among language levels; (2) In the *capito* guidelines, a

[5]Sentences taken from the brochure "Meine Rechte bei der Fähigkeits-orientierten Aktivität" published by the Austrian social government in 2011.

[6]In this analysis, the CEFR labels are not taken into account as corresponding to language levels.

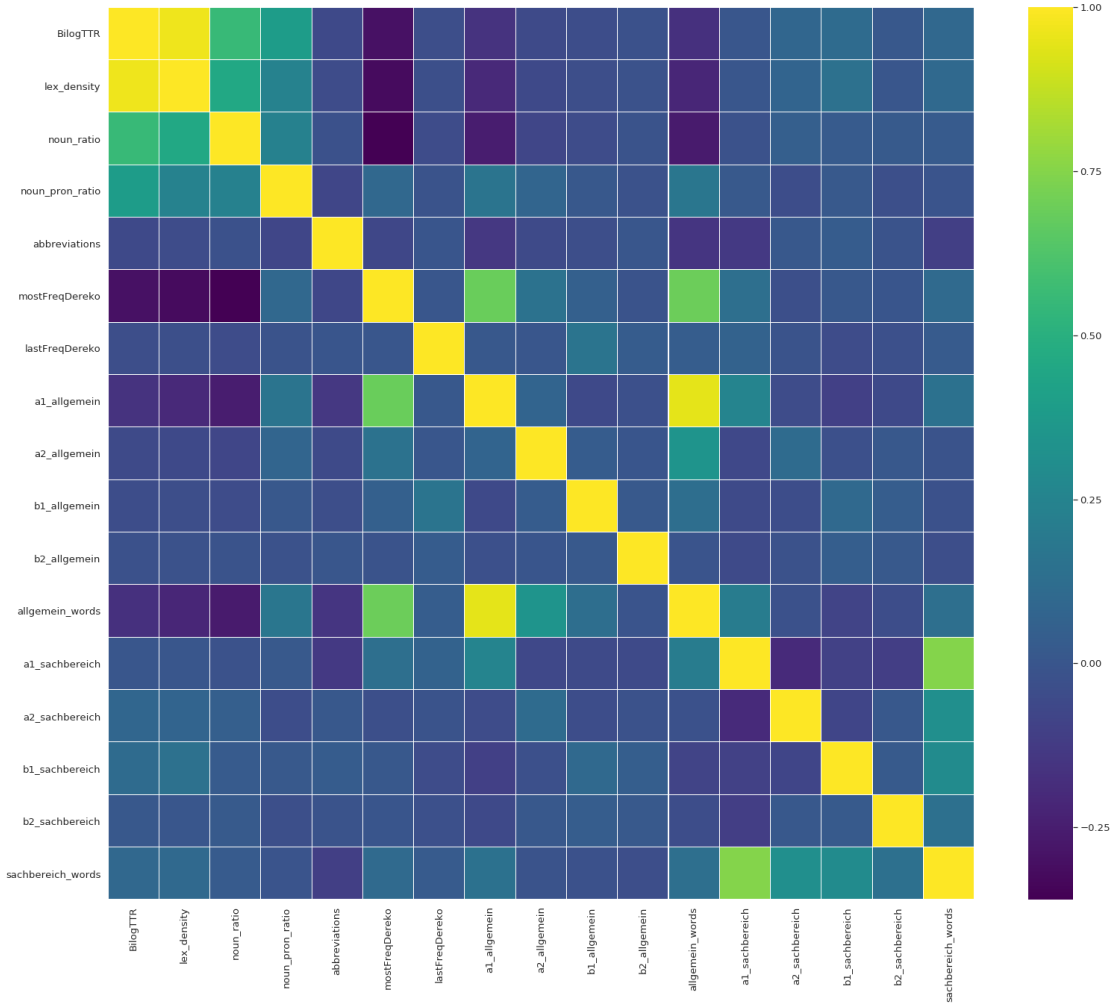
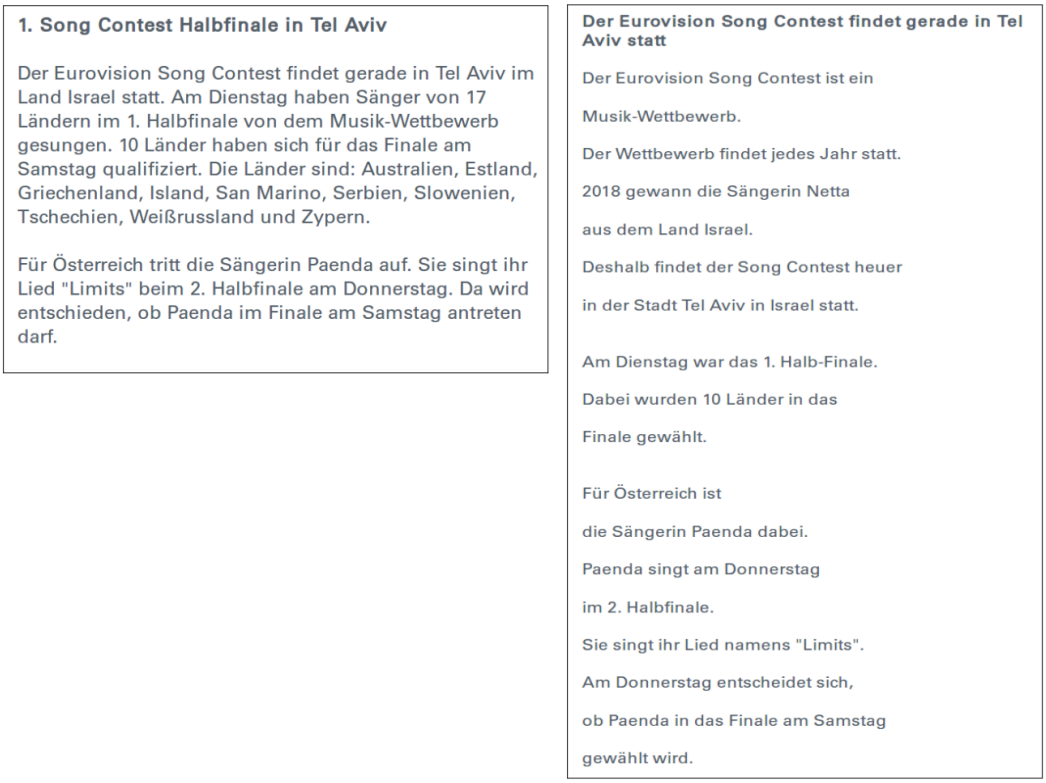


Figure 29: Plot of the correlation matrix of the lexico-semantic features (Subset 4) using the complete dataset and the Pearson correlation coefficient.

text at level B1 does not imply that every sentence has to be written on one single line (cf. Figure 30). If the second option is correct, as I can presume also by inspecting the HTML code of a APA webpage[7], then I can assume that this structural feature is fundamental in distinguishing among complexity levels. Future work might include further structural and typographic variables to explore their potential in distinguishing texts at different complexity levels.

As discussed in the results of each experiment, the feature agglomeration step led to a general improvement of the outcome (cf. Sections 3.4.2.2, 3.5.2.2 and 3.6.2.2). For this reason, it is useful to analyse the clusters of features in this context. The

[7] `https://science.apa.at/rubrik/bildung/Nachrichten_leicht_verstaendlich_vom_15_Mai_2019/SCI_20190515_SCI73194415448495606` (last accessed: 15 May 2019)



1. Song Contest Halbfinale in Tel Aviv

Der Eurovision Song Contest findet gerade in Tel Aviv im Land Israel statt. Am Dienstag haben Sänger von 17 Ländern im 1. Halbfinale von dem Musik-Wettbewerb gesungen. 10 Länder haben sich für das Finale am Samstag qualifiziert. Die Länder sind: Australien, Estland, Griechenland, Island, San Marino, Serbien, Slowenien, Tschechien, Weißrussland und Zypern.

Für Österreich tritt die Sängerin Paenda auf. Sie singt ihr Lied "Limits" beim 2. Halbfinale am Donnerstag. Da wird entschieden, ob Paenda im Finale am Samstag antreten darf.

Der Eurovision Song Contest findet gerade in Tel Aviv statt

Der Eurovision Song Contest ist ein
Musik-Wettbewerb.
Der Wettbewerb findet jedes Jahr statt.
2018 gewann die Sängerin Netta
aus dem Land Israel.
Deshalb findet der Song Contest heuer
in der Stadt Tel Aviv in Israel statt.

Am Dienstag war das 1. Halb-Finale.
Dabei wurden 10 Länder in das
Finale gewählt.

Für Österreich ist
die Sängerin Paenda dabei.
Paenda singt am Donnerstag
im 2. Halbfinale.
Sie singt ihr Lied namens "Limits".
Am Donnerstag entscheidet sich,
ob Paenda in das Finale am Samstag
gewählt wird.

Figure 30: Example of a news article at a level B1 (left) alongside with its A2 counterpart (right), from `https://science.apa.at/` (last accessed: 30 May 2019).

feature agglomeration algorithm treats each feature as a single variable, not knowing that some features are related to each other. This means it groups features solely based on their similar behaviour.

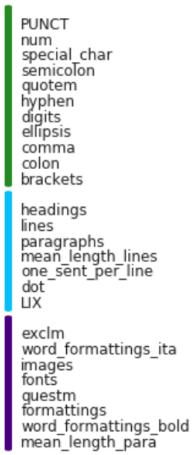


Figure 31: Plot of the agglomerated features of Subset 2.

Figure 31 shows three clusters for the surface features (Subset 2), which may correspond to distinctive aspects of this feature subset. The green cluster consists only of features at a character level, including punctuation and special characters. On the other hand, the cyan and the purple clusters contain features from the structural and typographic information, respectively.

Returning to the complete dataset, Figure 32a shows three clusters, as the results of the subexperiment presented in Section 3.4.2.2. The red cluster contains 32 features from various linguistic categories. Almost all surface features are included in this cluster. The orange cluster consists of 61 features mostly from the lexical, morphological and syntactic categories, while the yellow cluster includes 22 features from all sets. Specifically, it contains two features concerning auxiliary verbs in present and past tense (*VAUX*, *V_aux_praet*) and two features involving the genitive case (*gen_mod*, *gen*). Their clustering potentially indicates that these features are not discriminative enough.
Figure 32b concerns the feature agglomeration for the PDF subset. It shows the clustering results for the complete dataset after training the algorithm on 5 clusters. The cyan cluster may be an indicator of a wrong selection for the parameter `n_cluster`, since it contains only the feature for postpositions (*postpos*). All feature agglomeration experiments explain this assumption, showing better results with `n_cluster=3` (cf. Sections 3.4.2.2 and 3.5.2.2). What stands out from this figure is the orange cluster, which includes the feature for general words (*allgemein_words*) and general words at a simple level (*a1_allgemein*), verbs (*V*), full verbs (*VFULL*), finite verbs (*VFIN*), the most frequent words (*mostfreqDereko*), verbs at present tense (*V_pres*), main clauses (*main_clauses*) and full stops (*dot*). These linguistic characteristics are consistent with the most basic rules for simplified German (cf. Section 2.3.1). This indicates that these features do not behave differently, therefore they may be considered as one single variable. A correlation analysis might help to decide which features are redundant and might be congregated together. For instance, the correlation matrix in Figure 29 visibly shows that the variables *allgemein_words* and *a1_allgemein* are highly correlated (correlation coefficient between 0.75 and 1). The first feature, *allgemein_words*, may also be sufficient in the analysis and account for both variables.

In summary, linguistic and structural features account for multiple complexity levels. Future work would optimise and reduce the feature set based on further feature inspection, such as correlation analyses.

(a) Complete dataset (b) PDF files

Figure 32: Plot of the agglomerated features of Subset 1.

5 Conclusion

5.1 Conclusion

In this thesis, I have presented a clustering analysis on texts in simplified German in order to investigate evidence of multiple complexity levels.
I have given a literature review on automatic treatment of simplified texts, focusing on German. Research on automatic text simplification (ATS) and automatic readability assessment (ARA) has a long history in the literature, especially for the English language. Conversely, research on ATS and ARA on German and simplified German is relatively young. I have discussed linguistic properties that were used in previous studies to assess readability automatically. In this context, I have focused on those linguistic properties of simplified German that are relevant to distinguish simple from complex texts and emphasised the importance of structural and typographic features.

I have described the data I used for the clustering experiments, which included a parallel corpus of German/simplified German. This corpus consists of texts in simplified German, which were crawled from the web and aligned with their original German counterpart, if available. For my experiments, I considered only the simplified German versions. The corpus contains not only annotation at a linguistic level but also at a structural and typographic level. I have introduced a second text collection, which I prepared by applying the same preprocessing used for the first corpus to news articles made available by *capito* (cf. Section 1.1).

I have described the procedure of designing and automatically extracting features from corpora to use as input for the clustering algorithms. The features mostly corresponded to the linguistic properties identified during the literature review.
Subsequently, I have presented three clustering experiments. In the first, I have explored agglomerative hierarchical clustering, while in the second I have investigated the k-means partitioning. Both algorithms were run among five feature scenarios and different subsets of the simplified texts of the corpus of German/simplified German. In the third experiment, I replicated the approaches introduced for the first

and second experiments on the second text collection in order to investigate the clustering behaviour on a different dataset in simplified German. Moreover, I have examined the process of feature agglomeration, which reduces the feature space by clustering features instead of samples. This step led to the improvement of Calinski-Harabasz indices of on average 1500 pt. (between 15% and 60% in relation to the subset and algorithm analysed) in the results of all three experiments. This was also evident by visualising the dendrograms or scatter plots. Finally, I have explored the text clusters generated by the clustering algorithms and the feature clusters yielded by the feature agglomeration analysis.

The experimental approach with agglomerative hierarchical clustering outperformed the k-means approach on both datasets and managed to process the heterogeneous quality of the texts. The results confirmed a difference in complexity among texts. In other words, the texts were clustered so that the means of their corresponding linguistic features were distributed at different levels, depending on the analysed feature set. The results also highlighted the potential of structural features, such as the count of paragraphs and the rule *one-sentence-per-line*.

As this was the first attempt at investigating complexity in simplified German, there is no previous work that I could directly compare my approaches and results to. Since the analytical approach is based on machine learning, it can be applied to other languages in the future.
With this research, I shed light on the need to standardise simplified German and simplified languages in general. My work provides the basis for investigations of simplified-language varieties of other languages that do not depart from predefined levels of complexity.

5.2 Future Work

Simplified languages are low-resource languages, because the attention of institutions in e-accessibility has been growing only in the last decade or even in the last few years only. Machine learning algorithms ideally need more material than the 6,217 documents included in the corpus I used – this allows for learning efficient patterns and improving their performance. Therefore, the acquisition and preprocessing of texts in simplified language should be promoted, so that their analysis through machine learning techniques can produce more precise results. Scarton and Specia [2018] prove that ATS systems benefit from corpora built for different target groups, which make it possible to build better models than general-purpose ones. So, the next step for the corpus of German/simplified German (Section 3.1.1) is to check

if the predicted complexity levels correspond to the target users' needs and classify the texts into these levels to enhance the corpus annotation.

My experimental approaches did not take into account several additional aspects of texts, such as discourse or word senses. The subexperiments "webpages vs. PDFs" have demonstrated that features distribute differently according to their resource format (cf. Figure 28). Therefore, it would be interesting to take into consideration also the genre and domain specific nature of texts. The example of surface features has demonstrated that including structural and typographic features is beneficial to distinguish among complexity levels. For this reason, optical character recognition (OCR) systems could be exploited to include features corresponding to further structural information, such as number and width of columns, width of line spacing and presence of text indentations. Moreover, images as well as their captions and positional coordinates in a text could allow for investigations of image-text relationships or more advanced multimodal NLP tasks [Sanabria et al., 2018]. The type of image, as for example *photography* or *pictograms*, and its function in the text [Bock, 2018] can be also included in the surface feature set to enhance the discriminative power among different complexity levels of texts. The inclusion of knowledge-enriched word embeddings, "which encodes the knowledge on reading difficulty into the representation of words" [Jiang et al., 2018], could additionally bring a new perspective into the analysis and improve the overall performance.
Due to interpretability issues, I intentionally neglected neural network approaches or deep embedded clustering [Xie et al., 2015]. The next step can be, however, the exploration of more advanced techniques than agglomerative hierarchical clustering and comparison of the outputs.

With regard to the evaluation, clustering approaches in general need an appropriate evaluation method designed for resources without ground truth. The quality of my results has to be deeply investigated with such an approach that can carefully account for the final purpose of the analysis. The evaluation process might focus on the application of the methods and results on specific NLP tasks, such as ATS and lexical simplification, or consider the target audience while analysing simplified texts. Such an evaluation can additionally help to specify the relationship between needs of target users and definition of features describing multiple complexity levels.

Finally, the ideal scenario would be to exploit the analytical results to create a framework of inductively generated complexity levels and of linguistic properties included in these levels. This framework can feed into further research into and development of suitable ATS systems.

Glossary

corpus A collection of (machine-readable) texts.

feature A distinctive characteristic or property of a word or text that serves to distinguish it from other words or texts.

feature matrix An array of set of features that characterises a word or text distributed in rows and columns.

hyperparameter Configuration that cannot be directly estimated from the data by the model, such as `n_cluster` in k-means.

lemmatisation Process to assign the canonical base form, called *lemma*, to a word or phrase; e.g., *car* is the lemma of *cars*.

machine learning Computational technique to learn from existing data and predict unseen instances or modify process based on the acquired information.

machine translation Computational technique to translate spoken or written material from one source language to a target language.

parsing Process to analyse a string or a text into logical syntactic constituents.

POS-tagging Process of assigning a part of speech, i.e., a category to a word according to its syntactic function in a sentence.

simplified language Language variety characterised by reduced lexical and syntactic complexity; it includes images, structured layout and explanations for difficult words.

stopwords Words with little lexical meaning, such as function words, that are mostly omitted during NLP analyses.

tokenisation A linguistic preprocessing step that breaks a string or a text into individual units, called *tokens*.

vector An array of numbers. In a feature vector, the numbers correspond to the features of a lexical unit.

References

C. C. Aggarwal and C. X. Zhai. A Survey of Text Clustering Algorithms. *Mining Text Data*, pages 1–10, 2012.

S. M. Aluísio and C. Gasperin. Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, California, 2010. Association for Computational Linguistics.

M. Aranzabe, A. Ilarraza, and I. Gonzalez-Dios. First Approach to Automatic Text Simplification in Basque. In *Natural language processing for improving textual accessibility workshop programme*, pages 1–8, 2012.

B. Arfé, L. Mason, and I. Fajardo. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9):2191–2210, 2018.

D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, New Orleans, Louisiana, 2007.

G. Barlacchi and S. Tonelli. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2, pages 476–487, Samos, Greece, 2013.

A. Battisti and S. Ebling. A Corpus for Automatic Readability Assessment and Text Simplification of German. Submitted.

S. Bautista and H. Saggion. Making numerical information more accessible: The implementation of a Numerical Expression Simplification System for Spanish. *ITL - International Journal of Applied Linguistics*, 165(2):299–323, 2014.

B. Bengfort, R. Bilbro, and T. Ojeda. *Applied Text Analysis with Python*. O'Reilly, 2018.

A. Bernth. EasyEnglish: addressing Structural Ambiguity. In D. Farwell, L. Gerber, and E. Hovy, editors, *Machine Translation and the Information Soup*, pages 164–173, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

S. Bird and E. Loper. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, 2002.

C.-H. Björnsson. *Läsbarhet*. Liber, Stockholm, 1968.

B. M. Bock. "Leichte Sprache": Abgrenzung, Beschreibung und Problemstellungen aus Sicht der Linguistik. *Sprache barrierefrei gestalten*, pages 17–51, 2014.

B. M. Bock. "Leichte Sprache" - Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt. Technical report, Universität Leipzig, Institut für Förderpädagogik, Institut für Germanistik Institut für Sozialmedizin, Arbeitsmedizin und Public Health, 2018. URL `http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa2-319592`. (last accessed: 15 May 2019).

W. Bosma. Image retrieval supports multimedia authoring. In Z. S. E. and A. T., editors, *Linguistic Engineering meets Cognitive Engineering in Multimodal Systems, ITC-irst, ICMI Workshop*, pages 89–94, Trento, Italy, 2005.

U. Bredel and C. Maaß. *Leichte Sprache: theoretische Grundlagen, Orientierung für die Praxis*. Duden - Sprache im Blick. Dudenverlag, 2016.

L. Brouwers, D. Bernhard, A.-L. Ligozat, and T. Francois. Syntactic Sentence Simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, pages 47–56, Gothenburg, Sweden, 2014.

D. Brunato, F. Dell'Orletta, G. Venturi, and S. Montemagni. Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA, 2015. Association for Computational Linguistics.

D. Brunato, A. Cimino, F. Dell'Orletta, and G. Venturi. PaCCSS-IT: A Parallel Corpus of Complex-Simple Sentences for Automatic Text Simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas, 2016. Association for Computational Linguistics.

B. Bulté, L. Sevens, and V. Vandeghinste. Automating lexical simplification in Dutch Outline presentation. *Computational Linguistics in the Netherlands Journal*, 8:24–48, 2018.

Bundesministerium für Arbeit und Soziales. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung-BITV 2.0). Technical Report Teil 1, 2011. URL `http://www.bmas.de/SharedDocs/Downloads/DE/PDF-Publikationen/a712a-bitv-2.0-barrierefrei.pdf?{_}{_}blob=publicationFile`. (last accessed: 15 May 2019).

A. Candido, J. Sandra, and M. Aluísio. Building a Corpus-based Historical Portuguese Dictionary: Challenges and Opportunities. *TAL Traitement Automatique des Langues*, 50(2):73–102, 2009.

H. M. Caseli, T. F. Pereira, L. Specia, T. A. S. Pardo, C. Gasperin, and S. M. Aluisio. Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. In *10th Conference on Intelligent Text Processing and Computational Linguistics*, pages 59–70, Mexico City, Mexico, 2009.

M. Cha, Y. Gwon, and H. T. Kung. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006, Singapore, Singapore, 2017.

M. Cha, Y. L. Gwon, and H. T. Kung. Adversarial Learning of Semantic Relevance in Text to Image Synthesis. *Association for the Advancement of Artificial Intelligence*, 2019.

R. Chandrasekar, C. Doran, and B. Srinivas. Motivations and Methods for Text Simplification. In *Proceedings of COLING 1996, the 16th Conference on Computational Linguistics - Volume 2*, pages 1041–1044, Copenhagen, Denmark, 1996.

C. Chung and J. Pennebaker. The Psychological Functions of Function Words. *Social Communication*, pages 343–359, 2007.

K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Main Proceedings*, pages 193–200, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.

W. Coster and D. Kauchak. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon, 2011.

Council of Europe. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge, 2001. URL `https://rm.coe.int/1680459f97`. (last accessed: 15 May 2019).

P. Curto, N. Mamede, and J. Baptista. Automatic Text Difficulty Classifier - Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching. In *Proceedings of the 7th International Conference on Computer Supported Education*, volume 1, pages 36–44, Lisbon, Portugal, 2015.

R. M. DeKeyser. What makes learning second-language grammar difficult? A review of issues. *Language Learning*, pages 1–25, 2005.

F. Dell'Orletta, S. Montemagni, and G. Venturi. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK, 2011. Association for Computational Linguistics.

B. Drndarević, S. Štajner, S. Bott, S. Bautista, and H. Saggion. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2, pages 488–500, Samos, Greece, 2013.

W. H. DuBay. *The Principles of Readability*. Impact Information, 2004.

S. Evert, F. Dimpel, F. Jannidis, S. Pielström, I. Reger, C. Schöch, and T. Vitt. Statistical Significance in Literary Authorship Attribution. Technical report, Manchester, UK, 2019.

J. Falkenjack and A. Jonsson. Classifying easy-to-read texts without parsing. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, pages 114–122, Gothenburg, Sweden, 2014.

L. Feng. *Automatic Readability Assessment*. PhD thesis, The Graduate Center, City University of New York, 2010.

L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237, Athens, Greece, 2009.

W. Fleischer. *Wortbildung der deutschen Gegenwartssprache*. VEB Bibliographisches Institut, 1975.

T. François and C. Fairon. An "AI Readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea, 2012. Association for Computational Linguistics.

G. Freyhoff, G. Hess, L. Kerr, B. Tronbacke, and K. Van Der Veken. Make it Simple. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability. Technical report, ILSMH European Association, 1998.

C. Gasperin, E. Maziero, L. Specia, T. Pardo, and S. M. Aluisio. Natural language processing for social inclusion: A text simplification architecture for different literacy levels. In *XXXVI Seminário Integrado de Software e Hardware*, 2009.

M. Glaboniat, M. Müller, P. Rusch, H. Schmitz, and L. Wertenschlag. *Profile Deutsch*. Klett Langenscheidt, Berlin/Munich, Germany, 2005.

A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 2(36):193–202, 2004.

A. C. Graesser, D. S. McNamara, and J. M. Kulikowich. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 2011.

T. Grandin. *Thinking in pictures: And other reports from my life with autism*. Doubleday, New York, 1995.

M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

J. Hancke. Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. Master's thesis, University of Tübingen, Germany, 2013.

J. Hancke and D. Meurers. Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research Conference (LCR 2013)*, 2013.

J. Hancke, S. Vajjala, and D. Meurers. Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2102, the 24th International Conference on Computational Linguistics: Technical Paper*, pages 1063–1080, Mumbai, India, 2012.

K. Heimann Mühlenbock. *I see what you mean*. PhD thesis, University of Gothenburg, Sweden, 2013.

E. Hinrichs and S. Krauwer. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, 2014.

E. W. Hinrichs, M. Hinrichs, and T. Zastrow. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden, 2010. URL `http://www.aclweb.org/anthology/P10-4005`. (last accessed: 15 May 2019).

T. Hultman and M. Westman. *Gymnasistsvenska*. Skrifter / utgivna av Svensklärarföreningen. LiberLäromedel, 1977.

Inclusion Europe. Information für alle: Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht. Technical report, Inclusion Europe, 2009. URL `https://easy-to-read.eu/wp-content/uploads/2014/12/DE{_}Information{_}for{_}all.pdf`. (last accessed: 15 May 2019).

Institut für Deutsche Sprache. Korpusbasierte Wortformenliste DeReWo, DeReKo-2014-II-MainArchive-STT.100000. Technical report, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Germany, 2014. URL `http://www.ids-mannheim.de/derewo`. (last accessed: 15 May 2019).

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the Second International Workshop on Paraphrasing*, volume 16, pages 9–16, Sapporo, Japan, 2003.

M. Jansche, L. Feng, and M. Huenerfauth. Reading difficulty in adults with intellectual disabilities: analysis with a hierarchical latent trait model. In *Proceedings of the 12th International ACM SIGACCESS Conference on*

Computers and Accessibility, ASSETS '10, pages 277–278, Orlando, Florida, 2010.

Z. Jiang, Q. Gu, Y. Yin, and D. Chen. Enriching Word Embeddings with Domain Knowledge for Readability Assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378, Santa Fe, New Mexico, 2018.

T. Kajiwara and M. Komachi. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, 2016.

D. Kauchak, O. Mouradi, C. Pentoney, and G. Leroy. Text simplification tools: Using machine learning to discover features that identify difficult text. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2616–2625, 2014.

G. Kellermann. Leichte und Einfache Sprache – Versuch einer Definition. In *Aus Politik und Zeitgeschichte*, volume 64, pages 9–11. 2014.

D. Klaper, S. Ebling, and M. Volk. Building a German / Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings ofthe 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

S. Klerke and A. Søgaard. DSim, a Danish parallel corpus for text simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 4015–4018, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007. Association for Computational Linguistics.

T. P. Lau. Chinese Readability Analysis and its Applications on the Internet. Master's thesis, The Chinese University of Hong Kong, Hong Kong, 2006.

Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of Internal Clustering Validation Measures. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 911–916, Sidney, Australia, 2010. IEEE Computer Society.

H. Lüngen. DEREKO - Das Deutsche Referenzkorpus. *Zeitschrift fur Germanistische Linguistik*, 2017.

C. Maaß. *Leichte Sprache: Das Regelbuch*. Barrierefreie Kommunikation. Lit Verlag, 2015.

D. S. McNamara, A. C. Graesser, and M. M. Louwerse. Sources of text difficulty: Across the ages and genres. *Measuring up: Advances in how we assess reading ability*, pages 89–116, 2012.

H. Moisl. *Cluster Analysis for Corpus Linguistics*. De Gruyter Mouton, Berlin/Munich/Boston, 2015.

D. Naber. A Rule-Based Style and Grammar Checker. Diploma Thesis, Universität Bielefeld, Germany, 2003.

L. Naigles, M. Cheng, N. Xu Rattanasone, S. Tek, N. Khetrapal, D. Fein, and K. Demuth. "you're telling me!" the prevalence and predictors of pronoun reversals in children with autism spectrum disorders and typical development. *Research in Autism Spectrum Disorders*, 27:11–20, 2016.

Netzwerk Leichte Sprache. Die Regeln für Leichte Sprache. Technical report, 2013. URL `http://www.leichte-sprache.de/dokumente/upload/21dba{_}regeln{_}fuer{_}leichte{_}sprache.pdf`. (last accessed: 15 May 2019).

S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada, 2017. Association for Computational Linguistics.

P. Notter, C. Arnold, E. von Erlach, and P. Hertig. Lesen und Rechnen im Alltag. Grundkompetenzen von Erwachsenen in der Schweiz Eidgenössisches. Nationaler Bericht zu der Erhebung Adult Literacy & Lifeskills Survey. Technical report, Bundesamt für Statistik (BFS), Neuchâtel, Switzerland, 2006.

S. E. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1), 2009.

I. Pilan and E. Volodina. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings ofthe Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico, 2018.

S. Raschka. *Python Machine Learning*. Packt Publishing Ltd., 2015.

U. R. Raval and C. Jani. Implementing & Improvisation of K-means Clustering Algorithm. *International Journal of Computer Science and Mobile Computing*, 55(5):191–203, 2016.

K. Rayner and S. A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 1986.

L. Rello, R. Baeza-Yates, and H. Saggion. The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 501–512, Berlin, Heidelberg, 2013.

E. Rennes and A. Jönsson. A Tool for Automatic Simplification of Swedish Texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 317–320, Vilnius, Lithuania, 2015.

R. Reynolds. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, California, 2016.

J. Rogel-Salazar. *Data Science and Analytics with Python*. Chapman and Hall/CRC, New York, 1st edition, 2017.

H. Saggion. *Automatic Text Simplification*. Morgan & Claypool Publishers, 2017.

H. Saggion, E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg. Text simplification in Simplext: Making texts more accessible. *Procesamiento del Lenguaje Natural*, 47:341–343, September 2011.

R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze. How2: A large-scale dataset for multimodal language understanding. In *Proceedings of the NeurIPS Workshop on Visually Grounded Interaction and Language*, 2018. URL `https://arxiv.org/pdf/1811.00347`. (last accessed: 15 May 2019).

Š. Sanja, D. Biljana, and S. Horacio. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification / Eliminación de frases y decisiones de división basadas en corpus para simplificación de textos en español. *Computación y Sistemas*, 17, 2013.

S. Saraçli, N. Doğan, and İ. Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013(1), 2013.

C. Scarton and L. Specia. Learning simplifications for specific target audiences. In *56th Annual Meeting of the Association for Computational Linguistics*, pages 712–718, Melbourne, Australia, 2018.

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, pages 44–49, Manchester, UK, 1994.

W. Schnotz. *An Integrated Model of Text and Picture Comprehension*, pages 72–103. Cambridge University Press, second edition, 2014.

S. E. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530, Ann Arbor, Michigan, 2005.

R. Sennrich and B. Kunz. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1063–1067, Reykjavik, Iceland, 2014. European Language Resources Association.

R. Sennrich, G. Schneider, M. Volk, and M. Warin. A new hybrid dependency parser for German. In *Proceedings of the Biennal GSCL Conference*, pages 115–124, Potsdam, 2009.

R. Sennrich, M. Volk, and G. Schneider. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 601–609, Hissar, Bulgaria, 2013.

V. Seretan. Acquisition of Syntactic Simplification Rules for French. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 4019–4026, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

M. Shardlow. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications. Special Issue on Natural Language Processing*, 2014.

A. Siddharthan. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC'02)*, page 64. IEEE Computer Society, 2002. URL `http://dl.acm.org/citation.cfm?id=788016.788727`. (last accessed: 15 May 2019).

A. Siddharthan. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, Nancy, France, 2011.

A. Siddharthan. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165:259–298, 2014.

S. Sieghart. Leserlichkeit von Schriften als Faktor zur Textverständlichkeit von Texten in leichter Sprache. Technical report, 2018. Study by order of *capito*. Paper in review.

S. Sohangir and D. Wang. Improved sqrt-cosine similarity measurement. *Journal of Big Data*, 4(1):25, 2017.

R. R. Sokal and F. J. Rohlf. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40, 1962.

D. Sonagara and S. Badheka. International Journal of Computer Science and Mobile Computing Comparison of Basic Clustering Algorithms. *International Journal of Computer Science and Mobile Computing*, 3(10):58–61, 2014.

L. Specia. Translating from complex to simplified sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, Porto Alegre, RS, Brazil, 2010.

L. Specia, S. Frank, K. Sima'an, and D. Elliott. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation, Shared Task Paper*, volume 2, pages 543–553, Berlin, Germany, 2016. Association for Computational Linguistics.

S. Stajner and S. Nisioi. A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 3026–3033, Miyazaki, Japan, 2018.

S. Štajner, M. Franco-Salvador, S. P. Ponzetto, P. Rosso, and H. Stuckenschmidt. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 97–102, Vancouver, Canada, 2017. Association for Computational Linguistics.

J. Suter. Rule-based Text Simplification for German. Bachelor's Thesis, University of Zurich, Switzerland, 2015.

J. Suter, S. Ebling, and M. Volk. Rule-based Automatic Text Simplification for German. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 279–287, Bochum, Germany, 2016.

J. Sweller. *Implications of Cognitive Load Theory for Multimedia Learning*, pages 19–30. Cambridge University Press, 2005.

J. Sweller, J. J. G. van Merrienboer, and F. G. W. C. Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296, 1998.

United Nations. Convention on the Rights of Persons with Disabilities and Optional Protocol, 2006. URL `https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf`. (last accessed: 15 May 2019).

S. Vajjala. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. PhD thesis, Eberhard Karls Universität Tübingen, Germany, 2015.

S. Vajjala and K. Lõo. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, volume 107, pages 113–127, Uppsala, Sweden, 2014.

S. Vajjala and D. Meurers. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th workshop on building educational applications using NLP*, pages 163–173, Montréal, Canada, 2012.

T. Wang, P. Chen, J. Rochford, and J. Qiang. Text simplification using neural machine translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4270–7271, Phoenix, Arizona, 2016.

J. Xie, R. B. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. *CoRR*, abs/1511.06335, 2015. URL `http://arxiv.org/abs/1511.06335`. (last accessed: 15 May 2019).

W. Xu, C. Callison-Burch, and C. Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

V. Yaneva, I. Temnikova, and R. Mitkov. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

X. Zhang and M. Lapata. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

Z. Zhu, D. Bernhard, and I. Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of COLING 2010, the 23rd International Conference on Computational Linguistics*, pages 1353–1361, Beijing, China, 2010. Coling 2010 Organizing Committee.

Curriculum Vitae

Personal Information

Alessia Battisti
Imbisbühlstrasse 115, 8049 Zürich
alessia.battisti@uzh.ch

Education

2017 – 2019	Master of Arts in Multilingual Text Analysis University of Zurich
2012 – 2015	Master of Arts in Language Sciences Ca' Foscari University of Venice
2009 – 2012	Bachelor of Arts in Modern and Contemporary Languages and Civilizations Ca' Foscari University of Venice

Relevant Professional Experiences

2018 – 2019	Student Research Assistant at the Institute of Computational Linguistics
2017	MediNotice Project at the Institute of Computational Linguistics
2014 – 2015	GI-Tutor Project at the Ca' Foscari University of Venice

Further Professional Experiences

2016 – now	HR-Administrator at Avenir AG

Publications and Presentations

Battisti, A., Ebling, S. A Corpus for Automatic Readability Assessment and Text Simplification of German. (Submitted)

Battisti, A., Delmonte, R., Paschke, P. (2016). Automatic Detection of German Genitive Structures with GI-Tutor. In *Tagungsband der 38. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, Constance, Germany, 364-365.

Delmonte, R., Battisti, A. (2015). GI-TUTOR: Grammar-Checking for Italian Students of German. Presentation at the *Language Teaching, Learning and Technology*, Leipzig, Germany.
url: `http://www.isca-speech.org/archive/ltlt_2015/lt15_001.html`

A Example of a TCF file

This appendix shows a sample of a TCF file contained in the corpus German/simplified German [Battisti and Ebling, submitted]. The reference to the file is in the metadata entries.

```
<?xml version='1.0' encoding='utf-8'?>
<D-Spin xmlns="http://www.dspin.de/data" version="0.4">
 <MetaData xmlns="http://www.dspin.de/data/metadata">
 ...
  <Resources>
    <ResourceProxyList>
      <ResourceProxy id="idm46282901214496">
        <ResourceType>Resource</ResourceType>
        <ResourceRef>https://www.linz.at/images/MariaD.pdf
        </ResourceRef>
      </ResourceProxy>
    </ResourceProxyList>
    <JournalFileProxyList/>
    <ResourceRelationList/>
  </Resources>
  <Components>
    <OLAC-DcmiTerms>
      <alternative/>
      <contributor>Verein Hazissa</contributor>
      <contributor>capito Oberösterreich</contributor>
      <contributor>Müller, Silke</contributor>
      <date dcterms-type="W3CDTF">2016</date>
      <format dcterms-type="IMT">application/pdf</format>
      <identifier dcterms-type="URI">https://www.linz.at/images/MariaD.pdf
      </identifier>
      <language>A2</language>
      <publisher>Frauenbüro der Stadt Linz</publisher>
      <publisher dcterms-type="URI">www.linz.at/frauen</publisher>
      <rights/>
```

```
      <source>mariad.tetml</source>
      <subject>A2</subject>
      <tableOfContents>Maria sagt es weiter Seite 7; Informationen zu
      sexueller Gewalt Seite 12; Adressen von
      Beratungs-Stellen Seite 17; Wörterbuch Seite 32
      </tableOfContents>
      <title xml:lang="de">Maria sagt es weiter...
      Ein Bilder-Lese-Buch über sexuelle Gewalt und Hilfe holen.</title>
      <type dcterms-type="DCMITyp">Text</type>
      <type dcterms-type="DCMITyp">StillImage</type>
      <type/>
   </OLAC-DcmiTerms>
 </Components>
 ...
</MetaData>
<TextCorpus>
   <text>Vorwort Liebe Leserinnen! In diesem Heft geht es um ... </text>
   <tokens>
     <token ID="t_0" font="F0">Vorwort</token>
     <token ID="t_1" font="F1">Liebe</token>
     <token ID="t_2" font="F1">Leserinnen</token>
     ...
   </tokens>
   <sentences>
     <sentence ID="s_0" tokenIDs="t_0 t_1 t_2 t_3"/>
     ...
   </sentences>
   <textstructure>
     <textspan type="bold">Vorwort</textspan>
     <textspan type="paragraph" start="t_0" end="t_0"/>
     <textspan type="line" start="t_0" end="t_0"/>
     <textspan type="paragraph" start="t_1" end="t_3"/>
     ...
   </textstructure>
   <lemmas>
     <lemma ID="le_0">Vorwort</lemma>
     <lemma ID="le_1">lieb</lemma>
     ...
   </lemmas>
   <POStags tagset="stts">
     <tag ID="pt_0" tokenIDs="t_0">NN</tag>
```

```
      ...
    </POStags>
    <morphology>
      <analysis tokenIDs="t_0">Neut|Nom|Sg</analysis>
      ...
    </morphology>
    <depparsing>
      <parse ID="d_0">
        <dependency depID="t_0" func="root"/>
        ...
    </depparsing>
    <images>
        <image ID="I0" page="1" x="-1.07" y="112.47"/>
        ...
    </images>
    <fonts>
        <font id="F0" name="TradeGothic-BoldTwo"
        fullname="UDSPGZ+TradeGothic-BoldTwo" type="Type 1 CFF"
        embedded="true" ascender="977" capheight="722" italicangle="0"
        descender="-229" weight="700" xheight="520"/>
        ...
    </fonts>
 </TextCorpus>
</D-Spin>
```

B Topic Modelling Analysis

The process of topic modelling identifies the topics in a set of documents. For this analysis, I made use of the `textacy`[1] and `scikit-learn`[2] libraries to create non-negative matrix factorisation[3] (NNMF) and latent Dirichlet analysis[4] (LDA) models for the German/simplified German corpus (cf. Section 3.1.1).

To summarise, in NNMF, a normalised term-document matrix populated with tf-idf[5] values represents each text of the corpus. The matrix is then decomposed into two factors. The values of the factors is either positive or zero, while their product approximately corresponds to the original value in the matrix. This decomposition results in components that illustrate the topics, which are positively related to terms (i.e., words) and documents of the corpus or dataset.
The LDA technique represents the topics as the probability of words occurrence. Words occur in multiple topics and documents consist of various topics. This reflects the flexibility of the language. LDA estimates two matrices: The first includes the probability of choosing a specific word from a given topic; the second describes the probability of choosing a specific topic from a given document.

Prior to topic modelling analysis, I preprocessed the texts by deleting the stopwords. To extract the topics, I ran both models with different configurations, such as different number of topics (5, 10, 20). Table B1 shows the first six words for each resulting topic of the NNMF analysis. The topics can be guessed from the relations among words: Topic 1 refers to *webpages in simplified language*; topic 4 is about various *European countries and politics*; and topic 7 deals with *money*.

[1] `https://pypi.org/project/textacy/` (last accessed: 12 April 2019)

[2] `https://scikit-learn.org/stable/` (last accessed: 12 April 2019)

[3] `https://scikit-learn.org/stable/modules/decomposition.html#nmf` (last accessed: 12 April 2019)

[4] `https://scikit-learn.org/stable/modules/decomposition.html#latentdirichletallocation` (last accessed: 12 April 2019)

[5] This acronym stands for term frequency-inverse document frequency. Tf-idf is a statistical measure to evaluate the importance of words in relation to a document in a corpus.

Topic	Words					
0	menschen	beispiel	leben	arbeiten	bekommen	hilfe
1	leichter	sprache	internet·seite	informationen	seite	wahl·programm
2	wörter	gleiche	genaue	erklärung	beispiel	substantiv
3	wohnen	begleitete wohnen	begleitetes	pro	begleitperson	infirmis
4	land	deutschland	länder	russland	europa	frankreich
5	mann	frau	polizei	prozess	gericht	sex
6	sendung	radio	happy	nachfrager	hören	liachtblick
7	geld	bezahlen	bekommen	euro	staat	bekommt
8	rechte	steht	menschen	ausschuss	arbeits-gruppe	recht
9	behinderung	menschen	konvention	lebenshilfe	recht	inklusion

Table B1: Topics generated by the NNMF model with 10 topics.

Table B2 summarises the results generated by the LDA model with 10 topics and shows the first six words for each topic. Compared to the previous table, the topics are either more specific or too general. For example, topic 4 focuses on *technology and operation systems*, while topic 0 contains words from different domains that I cannot relate to each other intuitively (*Catalonia, bananas, partners, franchise, mushroom, gynaecology*).

Topic	Words					
0	katalonien	bananen	carles	partnerinnen	adile	franchise
1	gabi	tour	france	nsa	papst	klamm
2	fortbildung	sendung	flugzeug	nachfrager	drucker	bande
3	nachteils-ausgleiche	merk-zeichen	landes-blinden-fonds	landes-blinden-geld	grad	behinderung
4	iphone	betriebssystem	apple	linux	touchscreen	os
5	radio	happy	sendung	studio	special	olympics
6	kohl	uhren	helmut	kanzler·kandidat	kreuz·fahrt·schiff	sommerzeit
7	socken	fest·halten	flops	rektor	braun-kohle	hambacher
8	fähre	sylt	populismus	fähren	zug·strecken	schwäne
9	feier·tag	frauen-arzt	christi	walfänger	fach-arzt	pränatal

Table B2: Topics by the LDA model with 10 topics.

Both tables provide a generic overview of the topics included in the corpus. Further preprocessing steps, such as stemming and normalisation of words (e.g., by deleting the centred points or hyphens from compounds), may improve the performance of the models and lead to accurate topics. Additionally, hyperparameter tuning is necessary to obtain more precise results.

C Complete Feature Set

In the table below, I compare each feature in a hypothetical original text with the corresponding feature in its simplified counterpart. This functional distinction serves to deduce and operationalise the features that differentiate between simplified and original texts. All features are language independent, with the exception of features marked with *, which refers to the German language.

Table C1: Complete feature set.

Feature	**Simplified text**	**Original text**
Images	Number of images	No presence
Font type	One or two font types	Many different font types
Font size	Big	Small
Formatting devices (fd)	Single words with fd	Pieces of texts with fd
Number of columns	One	Multiple
Indentation	Visible	Sparse use
Line spacing	Big	Small
Text structure (lines, etc.)	Present	Lack of text structure
Paragraphs/Doc	High number	Low number
Sentences/Doc	High number	Low number
Lines/Doc	High number	Low number
Punctuation marks (pm)	High number of . and ?; Low number of other pm	Low number of . and ?; High number of other pm
Special characters	Low number	High number
Numerals	Arabic; digits (e.g., 4)	Roman numerals; Numbers in word (e.g., four)
Abbreviations and initial letters	Absent or with expansion	Present (without expansion)
LIX score	Low value	High value
Frequency word lists		
Named entities	Low number	High number
Nominal ratio	Low ratio	High ratio
Noun/Pronoun ratio	Low ratio	High ratio
Type/Token ratio	Low ratio	High ratio
Lexical density	Low density	High density
Lemma variation	Low variation	High variation
Negation	Absent or clearly marked	High use or not marked
Question words	High value	Low value
Number of proposition per sent.	Low number	High number
Entity density	Low density	High density
Lexical chains		
Word imagability		
Word concreteness	More concrete words	More abstract words
Word familiarity	More familiar words	Unfamiliar words
Ambiguity	Unambiguous words	Ambiguous words

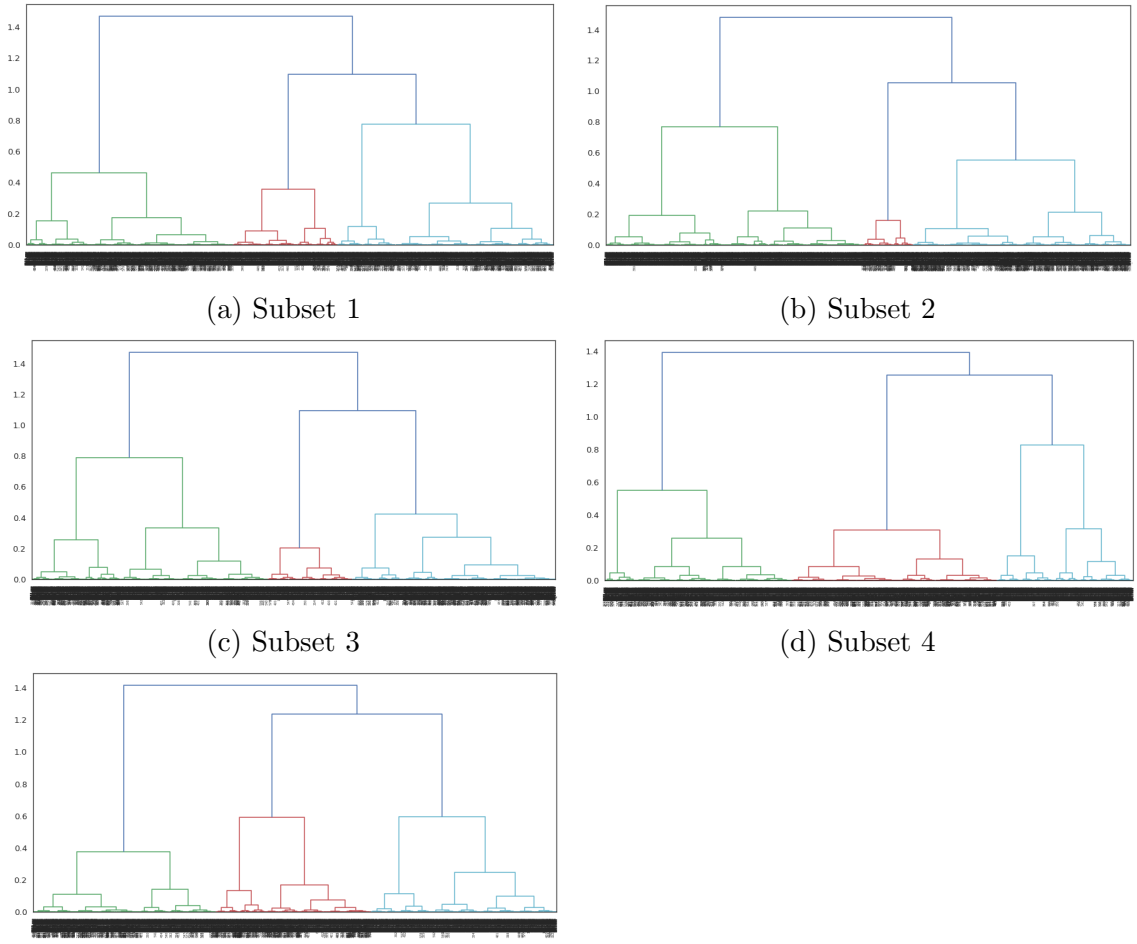
%

Table C1 – *Continued from previous page*

Feature	**Simplified text**	**Original text**
Word senses	One or low value	More senses per lemma
Attributive relation	*von*+dative construction	Genitive modifiers
Compounds	Compounds with - or $\cdot^*$	Compounds without - or $\cdot^*$
Comparative and superlative	Sparse use	Larger use
Verbal voice	Active	Passive (with or without agent)
Verbal mood	Indicative	Subjunctive, unreal, imperative
Verbal tense	Present	Simple past, past perfect, future
Count modal verbs		
Coordinating conjunctions	Low frequency	High frequency
Subjunctions	Low frequency	High frequency
Pre- and post-modifiers	Low frequency	High frequency
Adverbs	Local, temporal, prepositional	Other
Pronouns	Demonstrative; 1- and 2-pers. pron.	Personal pronouns; others
Particles	Low number	High number
Prepositions	Absent or sparse use	Prep. with genitives*
Standard word order	SVO*	SOV*
Sentence order	main clause - secondary clause	secondary clause - main clause
Number of main clauses	High number	Low number
Frequency of relative clauses	Low frequency	High frequency
Participial clauses	Low frequency	High frequency
Causal clauses	Low frequency	High frequency
Conditional clauses	Low frequency	High frequency
Concessive clauses	Low frequency	High frequency
Final clauses	Low frequency	High frequency
Interrogative clauses	Low frequency	High frequency
Modal clauses	Low frequency	High frequency
Temporal clauses	Low frequency	High frequency
Other clauses	Low frequency	High frequency
Parse tree depth clauses	Shallow	Deep
Avg. length of NPs, VPs, PPs	Shallow	Deep

D Dendrograms

Experiment 1: Dendrograms



(a) Subset 1

(b) Subset 2

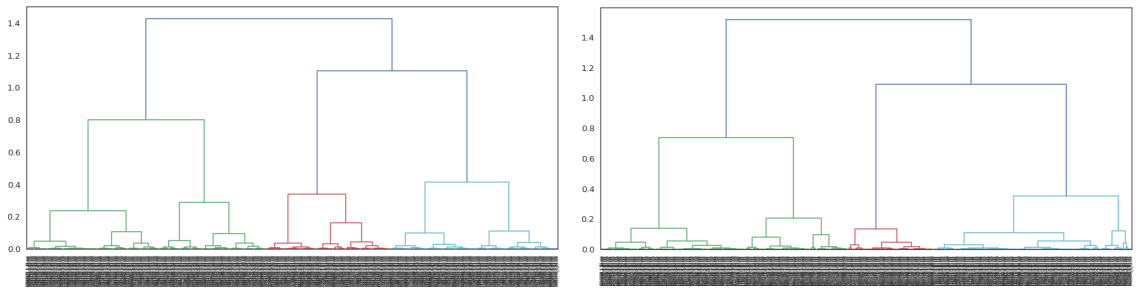
(c) Subset 3

(d) Subset 4

(e) Subset 5

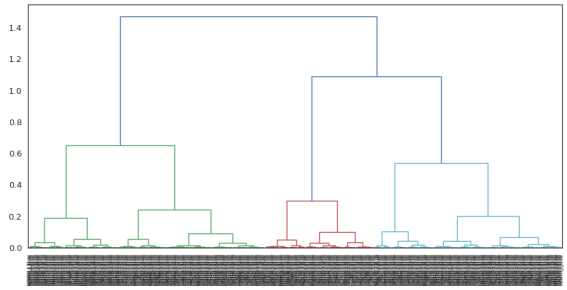
Figure 33: Visualisation of the dendrograms for the five feature subsets on the complete dataset.

Experiment 3: Dendrograms

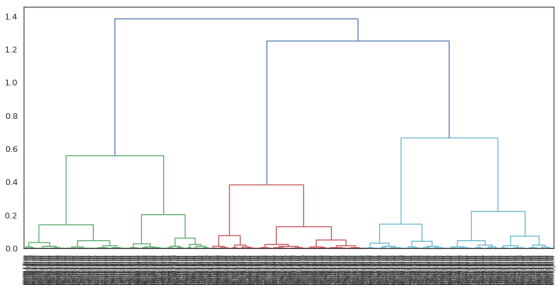


(a) Subset 1

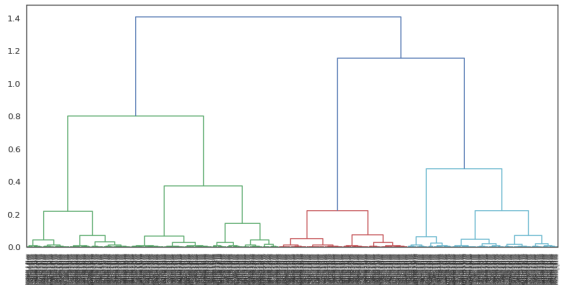
(b) Subset 2



(c) Subset 3



(d) Subset 4



(e) Subset 5

Figure 34: Visualisation of dendrograms for all five feature subsets on "TopEasy News" text collection.

E Tools and Resources

In the following, a list of tools and resources I used in this thesis.

Tools

- ParZu - The Zurich Dependency Parser for German
  code: `https://github.com/rsennrich/ParZu`
  demo: `https://pub.cl.uzh.ch/demo/parzu/`
- spaCy
  `https://spacy.io/`
- textacy
  `https://github.com/chartbeat-labs/textacy`
- TreeTagger
  `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`
- Zmorge
  `https://github.com/rsennrich/zmorge`

Resources

- German stopwords
  `https://github.com/solariz/german_stopwords`
- DeReWo-Grund-/Wortformenlisten
  `http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html`
- Profile Deutsch Allgemeine Begriffe
- Profile Deutsch Sachbereich

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass die Masterarbeit von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.uzh.ch/de/studies/teaching/plagiate.html).

Zürich, 13.06.2019 .....

Ort und Datum Unterschrift