Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstraße 19
72074 Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

B.A. Thesis in Computational Linguistics
(Version adjusted for distribution)

More Linguistically Motivated Features of Language Complexity in Readability Classification of German Textbooks: Implementation and Evaluation

Zarah Leonie Weiß

zweiss@sfs.uni-tuebingen.de

September 2015

First Examiner and Supervisor: Prof. Dr. Walt Detmar Meurers

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbstständig und nur mit den angegebenen Quellen und Hilfsmitteln einschließlich des WWW und anderer elektronischer Quellen angefertigt habe. Alle Stellen der Arbeit, die ich anderen Werken dem Wortlaut oder dem Sinne nach entnommen habe, sind kenntlich gemacht.

(Zarah Leonie Weiß)

Abstract

This thesis reports on the implementation of linguistically motivated features for the task of readabiliy classification of German. Five feature sets were designed, containing overall 46 features that are either a) linguistically more detailed re-implementations of common readability features, such as complex NPs and VPs, length measures of distances between Topological Field positions or inference markers of conditional clauses; or b) new features based on recent insights from research on the German academic writing register. These features were incorporated to the readability classifier by Hancke (2013) and Hancke, Vajjala & Meurers (2012) with and without the enhancements by Galasso (2014). Their performance was tested on the Reading Demands corpus consisting of texts from textbooks from German secondary schools by four different publishers. The classifier enhanced with the new features achieves an accuracy of up to 53.93% for grade level and up to 76.86% for school type classification and it improves classification compared to the non-enhanced classifier by up to 0.52% for grade- and 1.28% for school-wise classification. Also, the new features allow for detailed insights in the diverging linguistic properties of texts from different publishers.

Zusammenfassung

In dieser Bachelorarbeit werden neue, linguistisch motivierte Merkmale für die Leserlichkeitsklassifikation deutscher Texte vorgestellt. Insgesamt wurden 46 Merkmale erstellt und in fünf Merkmalsmengen gruppiert. Es handelt sich bei ihnen um a) bereits für Leserlichkeitsklassifikation verwandte Merkmale, die mit erhöhter linguistischer Detailfülle neu implementiert wurden, namentlich komplexe NPs und VPs, Längenmaße zwischen Positionen des Topologischen Modells oder Inferenzmarker von Konditionalsätzen. Zugleich wurden jedoch auch b) neue Merkmale implementiert, die auf rezenten Forschungsergebnissen zu den linguistischen Eigenheiten deutscher Wissenschaftssprache basieren. Die Merkmale wurden in den bereits bestehenden Leserlichkeitsklassifizierer aus Hancke (2013) und Hancke, Vajjala & Meurers (2012) eingebaut. Auch die Erweiterungen, die in Galasso (2014) berichtet werden, wurden in Teile des Klassifizierungsexperimentes inkorporiert. Getestet wurde auf dem Reading Demands Korpus, das aus Texten deutscher Geographiebücher des Gymnasiums und der Hauptschule besteht, die von vier verschiedenen Verlegern stammen. Die mit den neuen Merkmalen erweiterten Klassifizierer resultierten in Genauigkeitswerte von bis zu 53.93% für die Klassifizierung von Klassenstufen. Die Klassifizierung von Schultypen erreichte eine Genauigkeit von bis zu 76.86%. Im Vergleich zu den nicht erweiterten Klassifizierern konnte eine Genauigkeitszunahme von bis zu 0.52%, respektive 1.28% verzeichnet werden. Zudem gewähren die neuen Merkmale detaillierten Einblick in die unterschiedlichen linguistischen Strategien, mit denen die verschiedenen Verleger ihre Texte an die jeweiligen Rezeptionsbedürfnisse verschiedener Klassenstufen und Schultypen anpassen.

Acknowledgements

First, I would like to thank my supervisor Detmar Meurers for his great support and advice throughout the thesis. I am truly grateful for the effort and time he invested in this project.

I also thank Doreen Bryant for her valuable input and our discussions, without which half of this thesis would not have been possible. Sowmya Vajjala deserves my thanks for assisting me with my machine learning experiments.

Finally, I want to thank my parents Carsten and Heidi, my brother Max, Tanja and – of course – Jack. Without your sustained support and understanding this thesis would not have been possible, either (nor would have been the last one; you have my apologies for forgetting the acknowledgements there), and I, hereby, promise the five of you, that I will not write more B.A. theses. Probably.

Contents

List of Tables

List of Figures

1. Introduction

> "An average sentence, in a German newspaper, is a sublime and impressive curiosity; it occupies a quarter of a column; [. . .] it treats of fourteen or fifteen different subjects, each inclosed in a parenthesis of its own, [. . .] after which comes the VERB, and you find out for the first time what the man has been talking about; and after the verb – merely by way of ornament, as far as I can make out – the writer shovels in 'haben sind gewesen gehabt haben geworden sein,' or words to that effect, and the monument is finished. [. . .] I think that to learn to read and understand a German newspaper is a thing which must always remain an impossibility to a foreigner."
>
> *– Mark Twain (1880): The Awful German Language*

What impedes comprehension of language? Since the mediation of information is one important function of texts, the answer to this question is of relevance for authors as well as for text recipients; a text must match the reading skills of its audience. To identify readability automatically is one important task in the field of computational linguistics, but also subject of various other domains such as cognitive science, psychology, education, linguistics and Second Language Acquisition (SLA). Interestingly, some early, concrete and surprisingly rich suggestions can be found in Mark Twain's humorous elaborations on German newspaper texts quoted above: he names sentence length, idea density, verb position and verb clusters as elements contributing to the difficulty of newspaper texts from the perspective of a L2 speaker of German. Notwithstanding the satirical intentions of the essay *The Awful German Language*, the named characteristics agree with contemporary research on readability indicators, not only with respect to language learners, but to native speakers as well. It is noteworthy how heterogeneous those features are: while the length of text units is a highly superficial feature, that has already been employed in early readability formulas (e.g. Dale & Chall 1948), the assessment of idea density requires profound linguistic information. Also, while clause-final finite verbs and verb clusters are characteristics of only some languages, such as German, other aspects mentioned above hold without regard of the language: As the – generously shortened – first sentence in the quotation above artfully illustrates at the example of English, sentence length and idea density may be observed in various languages.

The field of readability classification investigates and implements features that are suited to indicate how challenging it is to read a specific text. As already mentioned, early approaches focused on superficial characteristics such as length. Recent research investigates the expressiveness of linguistically more ambitious features. Insights from related fields, such as SLA, have been employed with great success, considerably increasing classification accuracy, see for example Vajjala & Meurers (2012).

In the context of this development, this thesis examines the utility of five feature sets describing syntax and information organisation in German texts. More precisely, features were implemented describing the complex German NP and VP, Topological Field positions, deagentivation patterns and inference markers in conditional clauses. These disparate feature sets emerged in discussion with Detmar Meurers and Doreen Bryant. While some of the feature sets introduce new features that have been selected based on recent research discussions, others describe linguistically accurate and in depth features, that have already been captured superficially in previous approaches. In a classification experiment on German school textbooks, the feature sets proved to be overall beneficial for the task of readability classification and allowed to demonstrate remarkable differences in the adjustment of texts to different grade levels and school types between textbook publishers.

The thesis is structured as follows: the first section provides relevant background information, by briefly introducing work related to this thesis and the two main resources on which it is based: the readability classifier by (Hancke 2013; Hancke, Vajjala & Meurers 2012) and the Reading Demands corpus. Section 3 discusses each feature set with respect to its linguistic motivation, describes its empirical landscape and focusses on noteworthy aspects of the implementation. Afterwards, the classification experiment is reported in section 4: after a brief introduction to the general set-up, the performance of the feature sets and the rankings of the separate features in terms of information gain are reported. The section closes with a discussion of the data. Finally, section 5 concludes the results of the thesis.

2. Background

2.1. Related Work

The task of readability assessment has been addressed by researchers for over a century, as Vajjala (2015: 13) notes. It mainly addresses questions of text selection for educational purposes, but aside from this the readability of information material for adults, too, has been an ongoing domain of application.

The early approaches focused on easily accessible surface features of texts, due to the restricted possibilities of automatized in-depth analyses of language at that time. Vajjala (ibid.: 15) names Thorndike (1921) as the earliest approach to readability assessment, which focus on reading tasks for school children. In this context, see also Thorndike & Lorge (1944). Another early readability formula is the *Winnetka Readability formula* by Vogel & Washburne (1928). It considers the number of tokens, prepositions and uncommon words in a sample of 1,000 words from a text, as well as the number of simple sentences in a 75-sentence sample and is according to Vajjala (2015: 15) the first formula mapping readability to grade levels. However, the two most commonly known readability formulae are the *Dale-Chall* formula (Chall & Dale 1995; Dale & Chall 1948) and the *Flesch-Kincaid Grade Level* (Kincaid et al. 1975). Those two formulas mainly rely on sentence length measures and various word ratios, for example the ratio of rare words per text or of or personal words per text.

There has been criticism with regard to the validity of readability formulae, though, since they commonly access only superficial characteristics of texts and their language, see DuBay (2004, 2006) for a comprehensive overview. McNamara et al. (1996), for example, were able to produce increasingly difficult texts with a decreasing *Flesch-Kincaid Grade Level*. Henceforth, more recent approaches employ features based on more linguistic properties of difficult texts. They benefit from the increasing technical possibilities in computational linguistics and use insights from related areas, such as linguistics, SLA and cognitive science. Vajjala & Meurers (2012), for example, successfully employed features from SLA combined with traditional readability features to increase classification accuracy. Hancke, Vajjala & Meurers (2012) partially implement those features for German. For readability of French as a L2 language, François & Fairon (2012) use syntactic and semantic features as well as features specific to French as a foreign language. An other example is the *Coh-Metrix* project[1] (Crossley & McNamara 2011; Graesser et al. 2004), which uses insights from cognitive science for features measuring coherence and cohesion of texts. Some of those features were implemented for German by Galasso (2014). Other approaches employ language model features (Hancke 2013; Heilman et al. 2007) or discourse features (Feng & Jansche 2010; von der Brück 2007, 2008) for readability classification.

It is noteworthy, that von der Brück (2007, 2008), unlike most of the previously mentioned, performs readability classification not in an educational context. Instead, he tested his features on German administrative and municipal texts, referring to the task of inclusive information distribution: In order to make information accessible to an audience "with low literacy skills and/or with mild cognitive impairment" (cf. Dell'Ortella, Montemagni & Venturi 2011: 73) texts have to use so called *simple language*, which was defined for German in a guideline by the *Bundesministerium für Arbeit und Soziales*. Avoidance of passive, SVO word order, short sentences with a limited amount of words and the preferred usage of high frequency words are some of the characteristics of *simple language*. While Nietzio, Scheer & Bühler (2012) also investigate this special case of readability classification for German, Aluísio & Gasperin (2010) and Aluísio et al. (2010) do so for Portuguese newspaper texts within the *PorSimples* project and Dell'Ortella, Montemagni & Venturi (2011) for Italian newspaper texts with the *READ-IT* tool. Several studies evaluating the readability of medical information go in a similar direction, measuring the – considerable – discrepancy between text difficulty and audience literary. See for example D'Alessandro, Kingsley & Johnson-West (2001), Freda (2005), Gal & Prigat (2005), and Graber, Roller & Kaeble (1999). Those studies mostly employ superficial readability formulae, especially the above mentioned *Dale-Chall* formula and the *Flesch-Kincaid Grade Level*. It should, therefore, be noted that readability formulae are still commonly employed despite the ongoing criticism, see DuBay (2004) for a similar conclusion.

[1] http://cohmetrix.com.

2.2. The Readability Classifier

For experiment and implementation the German readability classifier from Hancke, Vajjala & Meurers (2012) was used as foundation including most extension from Hancke (2013), where the classifier was enhanced for German proficiency assessment. This original classifier models readability in terms of five feature groups of which four were employed for this thesis:

1. Features based on a language model (LM)
2. Features based on traditional readability formulae (Trad)
3. Features based on lexical information (Lex)
4. Features based on syntactic information (Syn)
5. Features based on morphological information (Morph)

The LM feature set was not replicated and excluded for the classification experiments. Also, in the most recent version of the classifier Trad features are incorporated in Lex and Syn and were not extracted into a separate feature group for the purposes of this thesis. Traditional readability features are in this case text length, average sentence length and word length in terms of characters and syllables. Lex, Syn and Morph contain several features known from the domains of readability classification, SLA, language proficiency assessment and language complexity, especially but not exclusively relying on work from Vajjala & Meurers (2012) and Lu (2010, 2011). For a more elaborated discussion of the separate features including the underlying formulas see Hancke (2013). Vajjala (2015), too, uses this classifier.

Galasso (2014) contributes features of text cohesion and discourse coherence to the original classifier leading to considerable increases in classification accuracy on the GEO-GEOlino Corpus. This corpus from Hancke, Vajjala & Meurers (2012) uses texts from the educational monthly *GEO*[2] magazine, which is written in German, and its special edition for children *GEOlino*[3]. The following feature sets were added to the classifier:

1. Features based on referring expressions (PrepDet)
2. Features based on referential indices (Ref)
3. Features based on connectives (Conn)
4. Features based on syntactic transitions (Tran)
5. Features based on simple text descriptions (Descr)

[2] http://www.geo.de.

[3] http://www.geo.de/GEOlino/.

They are mainly based on McNamara et al. (2014) and Todirascu et al. (2013) and were adapted to German for the purposes of the classifier. These feature sets were included in the classification experiment in section 4 as well.

Overall, the classifier consists of 175 features with and 135 features without the text cohesion and discourse coherence enhancement. The performance of both versions of the classifier within the experimental set-up used for this thesis is reported on in section 4, Table 4.1, p. 39.

The classifier as it was at disposal for this thesis employs several Natural Language Processing (NLP) tools for corpus preprocessing: For sentence segmentation and tokenization *SentenceDetectorME* and *TokenizerME* from OpenNLP[4], version 1.5.0 are employed, both using a publicly available pre-trained model for German from OpenNLP. Furthermore, the *Stanford Lexicalized Parser* was used to parse all texts, using the standard model for German, version 3.2.0 (Rafferty & Manning 2008), which is trained on the *NEGRA corpus*[5] (Skut et al. 1997). All texts were additionally parsed with the *Mate Dependency Parser* as well as the corresponding lemmatizer, version 3.6.0 (Bohnet 2010), again using the standard model for German (Seeker & Kuhn 2012) trained on the *TIGER corpus*[6] (Brants, Skut & Uszkoreit 2003). For a more detailed discussion of these tools please see Hancke (2013: 18–45). For the purposes of this thesis, the list of parsers was extended by the *Berkley Topological Field Parser* using the German model by Cheung & Penn (2009), which was trained on TüBa/DZ[7] (Telljohann et al. 2004). An example parse is given for the sentence *Ich habe den ganzen Kuchen alleine gegessen* (I ate the entire cake by myself) in Figure 2.1.

2.3. The Reading Demands Corpus

The Reading Demands corpus is a corpus of German Geography textbooks from which all corpora used for evaluations throughout chapters 3 and 4 were derived. The benefit of textbooks as corpora is that they stem from an actual domain of application for readability classification, see for example also François & Fairon (2012) and Vogel & Washburne (1928) for similar corpora for English and French. Vajjala (2015) uses the Reading Demands corpus, too, for her experiments on the original classifier, as discussed in section 2.2.

The version of the Reading Demands corpus which was used as starting material

[4] `https://opennlp.apache.org`.

[5] `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/`.

[6] `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.en.html`.

[7] `http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html`.

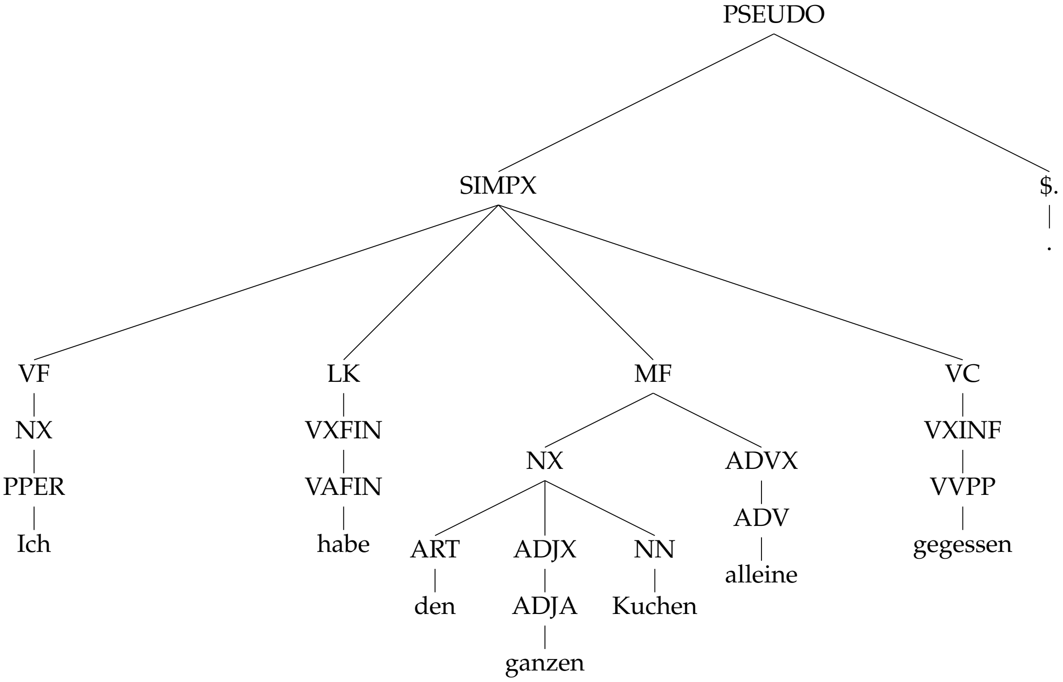

Figure 2.1.: Example output of *Berkeley Topological Field Parser*.

for this thesis consists of 2,928 texts from German textbooks used in schools in Baden-Württemberg state. It was crafted from 35 books from four different publishers using the OmniPage3 Optical Character Recognition software[8] (Vajjala 2015: 170). The texts range from 5^{th} to 10^{th} grade in either *Gymnasium* or *Hauptschule*, two types of secondary schools in Germany. While students acquire a higher education entrance qualification at the *Gymnasium*, German *Hauptschule* offers a lower secondary education. The texts within the corpus are evenly distributed along these two dimensions of school type and grade level. Since some texts were suited for two consecutive grades, for example 7^{th} as well as 8^{th} grade, the grade levels were grouped in doubles. For the classification experiment only 2,891 texts from the initial corpus were used; eleven files were corrupted and did not contain any data and 26 texts could not be processed due to technical issues during the preprocessing for syntactic transition counts. Galasso 2014: 17 reports similar issues in the GEO-GEOlino Corpus. This collection of 2,891 texts is henceforth referred to as the Full corpus and is displayed in Table 2.1.

Due to considerable differences between publishers, as reported in Vajjala (2015),

[8] `https://en.wikipedia.org/wiki/OmniPage`.

Grade level	School type		Total
	Gymnasium	Hauptschule	
5th to 6th	552	531	1083
7th to 8th	433	515	948
9th to 10th	454	406	860
Total	1,439	1,452	2,891

Table 2.1.: Full corpus of German Geography textbooks compiled from the Reading Demands corpus.

chapter 8, the Full corpus was further subdivided according to publishers into four sub corpora. Their structure is displayed in Table 2.2. The four Publisher A to D corpora differ in size ranging from 1026 to 333 texts. Also, while the Publisher A corpus shows a roughly even distribution of texts across the dimensions school type and grade level, the other three sub corpora are unbalanced: for example, the Publisher D corpus contains texts for *Hauptschule* only, which is evened out in the Full corpus since the Publisher B corpus and the Publisher C corpus contain more texts for *Gymnasium* than for *Hauptschule*.

A sixth corpus, the Reference corpus, was compiled consisting of ten randomly sampled texts from the Full corpus. It was used as a gold standard to measure the performance of the various features implemented to enhance the original classifier in terms of precision, recall and F-score on a domain specific corpus. This was necessary, to ensure the validity of the new features of which several were modelled with complex patterns. Only features achieving high performance scores were employed in the classification experiment in chapter 4. All texts in the Reference corpus were manually annotated by a single annotator with respect to each feature reported in chapter 3. Although multiple annotators and a larger set of texts would have been desirable to craft a gold standard neither was feasible in the course of this thesis. However, the texts were evenly distributed among grade levels and school types and showed most of the implemented features with sufficient frequency.

Corpus	Grade level	School type		Total
		Gymnasium	Hauptschule	
Publisher A	5th to 6th	238	156	394
	7th to 8th	141	223	364
	9th to 10th	116	152	268
	Total	495	531	1026
Publisher B	5th to 6th	114	126	240
	7th to 8th	145	70	215
	9th to 10th	108	59	167
	Total	367	255	622
Publisher C	5th to 6th	201	135	336
	7th to 8th	147	58	205
	9th to 10th	227	140	370
	Total	579	333	911
Publisher D	5th to 6th	–	114	114
	7th to 8th	–	164	164
	9th to 10th	–	55	55
	Total	–	333	333

Table 2.2.: Publisher A to D corpora compiled from the Reading Demands corpus.

3. Features

Five feature sets were modelled with regard to the dimensions of elaborateness and variance: Two feature sets describe syntactic phrasal complexity, namely the complex NP feature set and what is in analogy being referred to as the complex VP set. Phrasal complexity is a known feature from the field of language complexity, see Biber & Gray (2010), Lu (2010), and Lu & Ai (2015) and has already been included in work on readability classification, specifically in the original classifier from Hancke (2013). However, in this previous work phrase complexity as such was only measures in terms of overall number of modifiers and phrase length. Desideratum of this thesis was to implement a minutely detailed model of German complex NPs and VPs and, thereby, to investigate which aspects of complex phrases contribute most to the task of readability classification. The third feature set models positional properties of certain verbal dependants and their head in terms of topological field positions. Afterwards, a feature set is presented that collects deagentivational patterns as they are discussed in contemporary research on German academic language. The last feature set models conditional sentences with varying degrees of mediation of premise and effect. Those last three feature sets were implemented at suggestion of Doreen Bryant. In the following subsections each feature group is briefly motivated from a linguistic point of view. Also, conceptual decisions in the implementation will be critically reviewed. The performance of most implemented features was evaluated in terms of precision, recall and F-score for their counts based on the Reference corpus, albeit not all features could be tested with a suitable amount of instances. The full evaluation including the actual frequencies of true negatives, true positives and false negatives is displayed in Table B.1 on page 69 in appendix B, following the suggestion of Leacock et al. (2014: 35) to report those counts, too, whenever possible. If the performance of a feature count was insufficient or the number of occurrences too low to be decisive, it is pointed out in the discussion.

3.1. Features Based on Complex Noun Phrases

In its broadest definition a complex NP or nominal group is a NP that is extended by at least one dependant (see e.g. Gallmann & Lindauer 1994, Fabricius-Hansen 2014 and den Dikken & Singhapreecha 2004). This includes not only attributes in a narrower sense, which are adjuncts, but also arguments of NPs. Determiners, too, are part of complex NPs, as they are considered either NP-internal constituents, i.e. dependants, or heads of a determiner phrase (DP) with the NP as their complement, in which case what was called complex NP so far, actually is a complex DP.[1]

Complex NPs are an important feature of language complexity: the enhanced information density established by noun modifications as well as the increased ambiguity rate complicate comprehensibility (Schlömer 2013: 2), which has been confirmed in reading time studies, too, see DuBay (2004: 51). However, they also increase economy of texts and are a well known characteristic of German academic language, often referred to as as *Nominalstil* (Schlömer 2013: 1f, Henning & Niemann 2013: 447). Schlotthauer (2006: 1) names here especially noun modifying prepositional phrases (PPs), which are prominent in German complex NPs, but rare in most other European languages.

NP modification is not a new feature feature to readability classification. Graesser et al. (2004: 198) measure what they call *NP density* by computing a modifier ratio for both, NPs and VPs. A similar approach is employed in Dell'Ortella, Montemagni & Venturi (2011), Vajjala & Meurers (2012), and von der Brück (2008). Hancke (2013: 40) employs the same measure as Graesser et al. (2004) in the original classifier. However, the modifier ratio misses further linguistic details concerning the Part of Speechs (POSs) and positions of the modifiers. Yet, there is evidence that a more elaborate distinction of classifiers is desirable. Schlömer (Fig. 5 2013: 8) shows in her corpus study on German student's texts and school books that not only the amount but also the types of modifiers differ between academic and fictional texts as well as between grade levels (Graesser et al. 2004: 66ff). There are also early attempts to weight dependants differently based on their complexity. For example, Botel & Granowsky (1972) classify appositions, comparatives and dependant clauses as "3-count structures" in their readability formula and other modifiers as "2-count structures".

A weighting of the different dependants was not implemented to the classifier at

[1] This thesis remains theory neutral in this matter, but continues to use the term complex NP for sake of simplicity.

Nr.	Determiner	Pren. Modifier	Pren. Apposition	Head	Postn. Apposition	Postn. Modifier	Loose Apposition
1.	alle diese *all these*	neuen, ungewohnten *new, unfamiliar*	∅	Bräuche *customs*	∅	aus Übersee *from overseas*	∅ ∅
2.	keine *no*	besonders große *particularly strong*	∅	Abneigung *aversion*	∅ ∅	gegen Käse *against cheese*	∅
3.	Heidis *Heidi's*	frisch gebackenes *freshly baked*	∅	Brot *bread*	∅	dort auf dem Tisch *there on the table*	∅
4.	die beiden *the both*	∅	∅	Lieblinge *favourites*	∅	(von) meiner Mutter *of my mother*	, Max und Jack *, Max and Jack*
5.	welcher meiner *which of my*	∅	∅	Eintöpfe *stew*	∅	indisch / überbacken *Indian / gratinated*	∅
6.	∅	∅	∅	mir *(to) me*	∅	, als Linguistin, *as a linguist*	∅
7.	ein *a*	munterer *cheerful*	∅	Terrier *terrier*	∅	, wie man ihn kennt *as everybody knows them*	∅ ∅
8.	deines *your*	∅	∅	Onkels *uncle's*	Carsten *Carsten*	, den wir sehr lieb haben *who we love very much*	∅
9.	∅	∅	Onkel *uncle*	Carstens *Carsten*	∅	∅	∅

Table 3.1.: Positions in the German complex NP.

this stage, as it would have been beyond the possibilities of the thesis' time frame to determine reasonable weights. This issue may be addressed in future work. However, the implementation is designed to allow for an easy addition of such weights, for example attributive participles were counted separately from attributive adjective phrase (AP)s. This conceptual decision will be motivated further in a few paragraphs.

To capture all features related to complex NPs, first the empirical landscape was defined. As leading authority on the surface description of German grammar, Duden (Gr) (2009)[2] (§ 276-292) was consulted to build a comprehensive model of the parts of speech of noun dependants, their positions in the NP and their characteristics. The resulting map of the complex German NP is illustrated in Table 3.1: Located between determiner and the head of the NP are prenominal modifiers. Those are attributive adjectives and participles (rows 1-2, 7). They agree with determiner and head in gender, case, number and definiteness. They are recursively iterable and can themselves be modified, for example by adverbs or adjectives (row 1-3).

Postnominal modifiers are located to the right of the head noun and may also branch and be recursively iterated (row 3). However, they are not inflected. Common phrasal postnominal modifiers are PPs, which may be adjuncts or arguments of the head. Other postnominal modifiers are adverbs and Genitive attributes which may be either NPs or PPs (row 1-4). However, in rare cases they may also be

[2] I.e. the *Duden **Gr**ammatik* (Duden Grammar) of German.

postponed participles or adjectives (row 5), for example in gastronomical language, cf. Duden (Gr) 2009, §277: *Forelle blau* (lit.: trout blue). Other types of postnominal modifiers are comparative groups, which may either be phrasal and introduced by *als* or clausal and introduced by *wie* (row 6, 7). Finally, relative clauses, too, count as postnominal modifiers (row 8).

The only constituents that can be inserted between pre- and postnominal modifiers and the head noun are close appositives. They are typically NPs themselves, e.g. titles or kinship terms combined with a proper noun or forename and surname (row 8, 9). Postnominal close appositives always co-occur with determiners, while prenominal close appositives occur without determiners. This shift is illustrated with the Genitive case marking displayed in both rows. As in German only the head noun but not the appositive inflects according to the phrase's case, the noun bearing the Genitive case marking *-s* has to be the head noun. Unlike close appositives, loose appositives (row 4), are usually separated by commas and occur at the rightmost periphery of the NP.

Determiners are the elements occurring at the leftmost periphery of most NPs containing a singular head. They are definite and indefinite articles as well as attributive possessive, demonstrative, interrogative and indefinite pronouns. Prenominal NPs in Genitive case are located in the determiner position, too. Although each NP has usually at most one determiner, in some restricted cases combination of two adjacent determiners containing only one attributive pronoun are possible (row 1, 4, 5), cf. ibid., §302.

As a last remark, it as to be stated that in German heads of complex NPs are typically but not obligatorily nouns. This is illustrated by row 6, where a personal pronoun is the head of a NP extended by a comparative dependant. Although rare and incompatible with most extensions of complex NPs, these cases have to be taken into consideration, too.

Based on this empirical template a feature set consisting of nine features was implemented, in the following referred to as CompNP. The features and their respective computation formulae are listed in Table 3.2. While the first eight features measure the elaborateness of complex NPs, the last feature was designed to capture the variance of noun modification in terms of the ratio of observed to possible noun phrases.

In order to conceptualise the displayed dependants of complex NPs, two approaches were employed: Appositions / parentheses and noun modifying relative clauses were identified head-wise using dependency labels provided by the *Mate*

Feature	Formula
Ratio of determiners	# determiners / # NPs
Ratio of possessive noun attributes	# possessives / # NPs
Ratio of prenominal attributive adjectives or adverbs	# prenominal adjectival or adverbial modifiers / # NPs
Ratio of postnominal noun modifiers	# postnominal modifiers / # NPs
Ratio of attributive participles	# participle I or II attributes / # NPs
Ratio of appositions or parentheses	# appositions or parentheses / # NPs
Ratio of comparative noun modifiers	# comparative noun modifiers / # NPs
Ratio of clausal noun modifiers	# clausal noun modifiers / # NPs
Coverage of noun modifications	# observed noun modifier types / # possible noun modifiers

Table 3.2.: Features based on the German complex NP.

Dependency Parser. However, the parser labelled appositions / parentheses with recall of 0.222. With precision 1.000 this lead to a F-score of 0.363, which was considered insufficient: ratio of appositions / parentheses was excluded from the feature set. Possessive noun modifiers, which include pre- and postnominal NPs in Genitive case as well as possessive pronouns, were identified similarly, additionally using case information also provided by the *Mate* parser. The feature was introduced, and the possessives thereby separated from other determiners, prenominal modifiers or postnominal modifiers, because from a variationalistic perspective it is often argued that possessive noun modifiers are forms of the same linguistic item, see e.g. Payne & Berlage (2014), Rosenbach (2014), and Wolfram (2006). It seemed, therefore, reasonable to group them separate from other modifiers.

The other dependants were identified sentence-wise via *Tregex*[3], which is a pattern matching utility for trees by Levy & Andrew (2006) from the Stanford NLP group. Five *Tregex* patterns were designed based on the parses provided by the *Stanford* parser. The patterns search for specific POS tags within the c-command domain of a noun. This includes determiners in Stanford parse trees, due to their n-ary branching, which is illustrated in Figure 3.1 for the complex NP *Die Bewegung der Erde um die Sonne* (the movement of the earth around the sun), which is an actual example from the Reference corpus. The c-command domain was marked with red. In the case of pre- and postnominal modifiers information on the linear order of modifier and head noun was employed. It should be noted that the dependants of complex AP modifiers were treated as dependants of the dominating NP, in order to capture the increased complexity of an attributive complex AP compared to a simple AP. Therefore, the corresponding feature was called ratio of prenominal attributive adjectives or adverbs, although there are no prenominal attributive adverbs as such in German.

[3] http://nlp.stanford.edu/software/tregex.shtml.

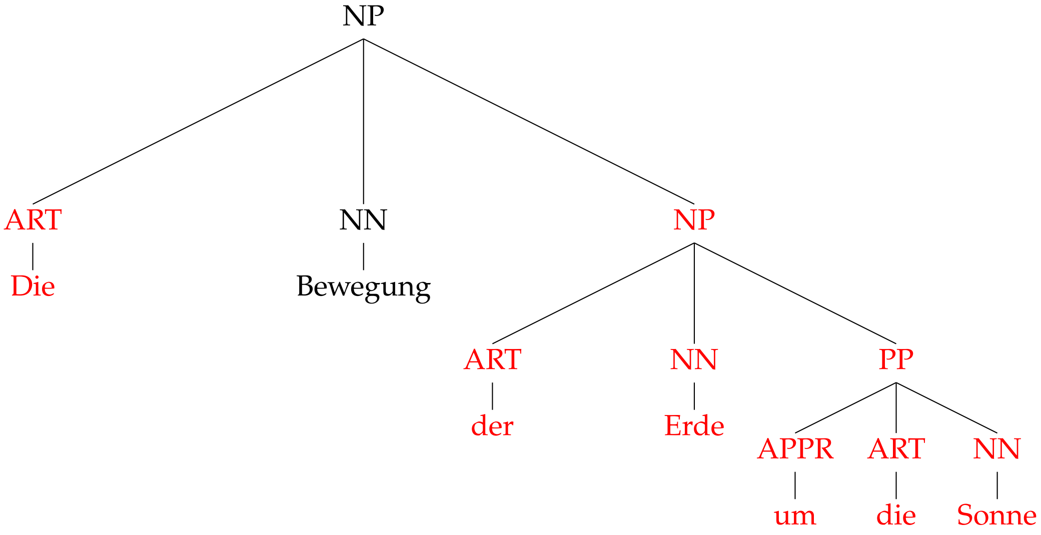

Figure 3.1.: C-commando Domain of the head noun *Bewegung* in a Stanford constituency parse.

While the domain of prenominal attributive modifiers was artificially broadened, it was also shrinked by the heuristic distinction between adjectival or adverbial prenominal modifiers and attributive participles. As already briefly mentioned, in terms of formal semantics, verbs, participles included, denote events or states[4] and have, therefore, an ontologically different and more complex status than adjectives (Zimmermann 1999: 125). It is, therefore, desirable to differentiate between both types of prenominal modifiers. However, the *Stanford* parser tags prenominal participles as adjectives. Although strictly speaking incorrect, this is a reasonable decision for a NLP tool: The distinction is most likely irrelevant in most cases, because in German attributive participles serve the same purpose as adjectives. Furthermore, to differentiate them is not a trivial task, at least not for past participle due to its various forms. While weak and mixed verbs build the past participle consistently with the prefix *ge-* and the suffix *-(e)t*, strong verbs commonly build their past participle with the prefix *ge-* and the suffix *-en*.[5] Yet, some verbs build their past participle without the prefix *ge-*, namely all complex verbs starting with *er-*, *ver-*, *zer-*, *be-*, *ge-*, *ent-*, *emp-* or *hinter-*, and all verbs ending in *-ieren*. Also, all

[4] For an introduction to event semantics, see Davidson (cf. 1967) and Vendler (1967).

[5] In German weak verbs retain their root vowel in all tenses, for example *lachen* (to laugh) – *lachte* (laughed) and show regular inflection. In contrast, strong verbs change their root vowel from simple present to simple past, for example *laufen* (to walk) – *lief* (walked), and tend to show irregular inflection. If they inflect regularly despite of the vowel change, they are called mixed verbs.

prefix verbs building the past participle with *ge-* insert it between prefix and verb stem. The *Tregex* pattern included all these cases, except for prefix verbs whose prefix is not one of those listed above. Naturally, the plurality of patterns increased the error susceptibility. For example, the wider pattern includes several written out enumerations, i.e. *ersten* (first), *zweiten* (second) ... Also, some adjectives are polysems to participles, such as *gedacht* (either *imaginary* or past participle of *to think*), or homonyms, such as *verschieden* (either *different* or past participle of *to pass away*). These examples origin from actual errors made on the Reference corpus. Overall, the pattern lead to precision, recall and F-Score of 0.840, which leaves room for improvement, but was considered sufficient to retain the feature. Opposed to past participle, present participle can be identified relatively effortlessly by its suffix *-end*, i.e. the infinitival suffix *-en* plus the present participle suffix *-d*, possibly followed by German adjectival inflection. This definition is robust as a pattern, because it applies without exception to all present participle forms and at the same time excludes all adjectives, which – to my knowledge – never end in *-end*. Accordingly, attributive present participle counts reached 100% in precision, recall and F-score. Regarding this score one has to be aware that the number of attributive present participles in the gold standard was relatively low with only 4 instances. However, it is to be expected that all scores would remain high with higher numbers of occurrences.

3.2. Features Based on Complex Verb Phrases

Unlike complex NP, the notion of complex VPs is not a standing concept in linguistics. However, adjectives, adverbs, prepositions and participles modify not only NPs, but also VPs. To refer to these enhanced VPs, the term complex VP is used for the purposes of this thesis. The notion also includes other complexity increasing elements of VPs, namely verb clusters and verb participles. Complex VP is, therefore, unlike complex NP not a linguistic notion, that was realised in form of a feature set. Instead, the complex VP feature set, henceforth CompVP, collects linguistically heterogeneous features that indicate increasing complexity in a VP or the verb complex (VC). The VC is the position in a German clause where non-finite verbs and separated verb particles are located.

As noun modifiers, verb modifiers have been counted in Hancke (2013) already in a general fashion, as well as in other readability classifiers, for example Dell'Ortella, Montemagni & Venturi (2011). However, by the same line of reasoning as for

Prefield	C/FIN	Middle Field	VC
$[\text{Schnell}]_{Adj}$	$[\textbf{gehe}]_V$	ich $[\text{abends}]_{Adv}$	–
$[\text{Abends}]_{Adv}$	bin	ich $[\text{schnell}]_{Adj}$	$[\textbf{gegangen}]_V$
$[\text{Am Abend}]_{PP}$	$[\textbf{geht}]_V$	sie $[\text{einen Keks in die Tasche steckend}]_{PresPart}$	–
$[\text{Einen Keks in die Tasche gesteck}]_{PastPart}$	ist	sie $[\text{am Abend}]_{PP}$	$[\textbf{gegangen}]_V$

Table 3.3.: Verb modifiers in the (reduced) Topoligical Field model.

the previous feature set, it seemed fruitful to analysis verb modifying adjectives, adverbs, participles and PPs separately. They are similar to the corresponding noun modifiers, yet any verb modifier may precede or follow its head verb without regard of its POS, because the order of verb and modifier is determined by their positions in the Topological Field model, for an introduction to the Topological Field model, please see Höhle (1986):

Verb modifiers are located in the middle field, but may be fronted to the prefield. The modified verbs may be positioned in the C/FIN position, where either the complementizer or the finite verb of a clause may go, or in the VC. These two positions also known as left sentence bracket (LSB) and right sentence bracket (RSB) are and crucial for the structure of German clauses and the Topological Field model. Since the positions in the Topological Field are entirely independent from the actual modification process, the notion of pre- and postverbal modifiers is descriptive only for specific instances, but does not allow for any generalisations with regard to the verb-modifier relation. This is illustrated in Table 3.3, which shows a reduced version of the Topological Field model. Modifier and head verb constituents are marked with square brackets and flagged with their respective POS. The modified verb is additionally highlighted with bold font.

Additional to these modifiers, verb clusters are also grouped within this feature set. In verb clusters multiple adjacent verbs form a complex verbal phrase. Verb clusters are referred to as V2, V3, etc. according to their size, which is in principle unrestricted in German. Yet, cluster size and frequency are inversely correlated as shown in Figure 3.2. The plot displays verb cluster occurrences in DWDS[6], including the core corpus, the weekly actualised German text archive and the *Zeit* corpus. With increasing cluster size, verb clusters also become more difficult to process, which is illustrated with a V7 cluster in Example 1. This clause is highly artificial and even for native speakers only comprehensible after repeated thorough reading. Its construction, too, required a step-wise approach.

[6] `http://www.dwds.de`.

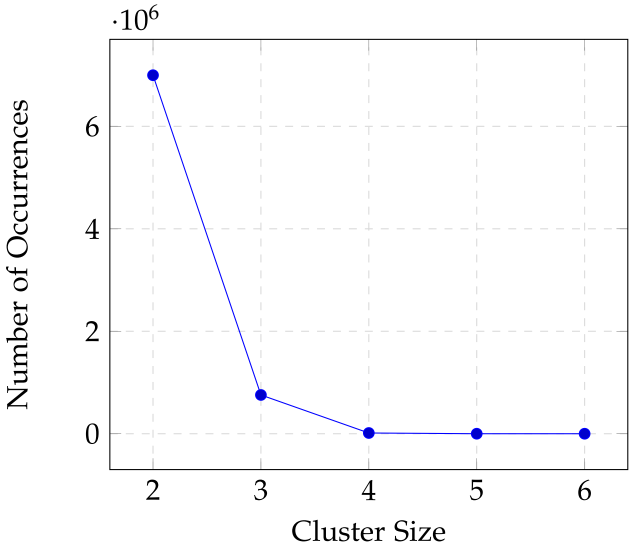

Figure 3.2.: Number of occurrences of V2 to V6 clusters in DWDS.

(1) dass du mir **einzuschlafen geholfen haben gewollt haben**
that you me.Dat fall-asleep.Zu help.PP have.Inf want.PP have.Inf
können wirst
can.Inf will.2.Sg.Pres
'that you will have been able to have wanted to have helped me to fall asleep'

The high amount of V2 clusters in Figure 3.2 can be explained by the fact that periphrastic grammatical constructions form V2 clusters by default in verb final clauses. Accordingly, auxiliary verbs induce verb clusters most commonly. On DWDS 5,125,330 instances of auxiliary verbs governing an adjacent verb were found. With 1,727,708 occurrences, modal verbs are distant second in terms of frequency. The rarest kind of verb clusters are those in which main verbs select each other as in *einzuschlafen helfen* (to help to fall asleep), occurring 756,665 times in DWDS. Periphrastic German tenses are present and past perfect and future 1 as well as future 2. Additionally, the passive voice is realised periphrastic in German. Example 2 shows a past perfect induced V2 cluster in active voice. If periphrastic tense and passive are combined, the cluster size is always increased by one, as shown in Example 3 for future 2.

(2) dass es **geregnet hatte**
that it rain.PP have.3.Sg.Past
'that it rained'

Feature	Formula
Ratio of adjectival / adverbial modifiers	# adjectival + adverbial modifiers / # VPs
Ratio of prepositional modifiers	# prepositional modifiers / # VPs
Ratio of participle modifiers	# past participle + # present participle / # VPs
Ratio of verb particles	# verb particles / # inflected verbs
Ratio of phi feature sub clusters	# phi feature sub clusters / # sub clusters
Ratio of modal verb sub clusters	# modal verb sub clusters / # sub clusters
Ratio of main verb sub clusters	# main verb sub clusters / # sub clusters
Average verb cluster size	# V2 + V3 + V4 + V5 + V6 / # verb clusters
Ratio of periphrastic tenses	# present perfect + past perfect + future 1 + future 2 / # finite verbs
Ratio of present perfect	# present perfect / # finite verbs
Ratio of past perfect	# past perfect / # finite verbs
Ratio of future 1	# future 1 / # finite verbs
Ratio of future 2	# future 2 / # finite verbs
Ratio of simple present	# simple present / # finite verbs
Ratio of simple past	# simple past / # finite verbs
Coverage tenses	# observed tenses / # possible tenses
Coverage periphrastic tenses	# observed periphrastic tenses / # possible periphrastic tenses
Coverage verb cluster sizes	# observed verb cluster sizes / # possible verb cluster sizes
Coverage sub cluster types	# observed sub cluster types / # possible sub cluster types
Coverage modifier types	# observed modifier types / # possible modifier types
Variance of cluster sizes	$\sum_{i=2}^{6}(i * \#Vi-$ average cluster size) / # verb clusters

Table 3.4.: Features based on the German complex VP.

(3) dass ich lieber vorher **gefragt worden wäre**
that I rather in-advance ask.PP get.PP be.1.Sg.Pres.Subj
'that I would have preferred to be asked in advance'

Finally, separated verb particles were also considered, as they postpone the interpretation of the main verb similar to verb clusters. This feature is also employed in other classifiers, for example by von der Brück (2008). Ideally, complex verbs in general would have been included to the feature set, i.e. participle and prefix verbs in contrast to simple verbs. Unfortunately, due to technical issues it was not possible to include a suitable morphological parser to the classifier in time for this thesis. The enhancement of the complex VP feature set by all types of complex verbs remains, therefore, for future work.

Table 3.4 shows the features implemented to capture the described components of enhanced VPs. Again, features were created to measure elaborateness and variance, the latter mostly in terms of coverage of possible patterns. However, for cluster size the actual variance was measured, too. As for elaborateness, two approaches were employed to collect the counts: Similar to the complex NP feature set, verb modifying adjectives, adverbs and PPs were identified using a combination of dependency labels and POS tags provided by the *Mate* parser. Since verb modifying past participles are in fact POS tagged as participles, they could be selected with this approach, too. However, present participle modifiers are tagged as adjectives

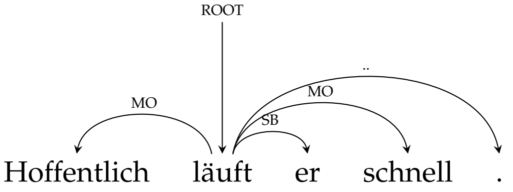

Figure 3.3.: *Mate* parse for sentence-level (*hoffentlich*) and verb-level (*schnell*) modification.

and were, therefore, identified as before as all adjectives ending in *-end*.

It should be pointed out that the *Mate* parser does not allow to differentiate between sentence- and verb-level modifications as illustrated in Figure 3.3 or 'frame-setting modifiers" and event modifiers: while the former impose an underspecified restriction on the entire proposition, the latter modify the situation referent of a verb (Maienborn 1999: 46), see also Maienborn (2001) for a detailed discussion. Due to the qualitative difference in the nature of modification a separate acquisition of these modifiers would be desirable and might be approached in future work, if a feasible identification method can be found.

Verb cluster related features were identified sentence-wise using elaborated *Tregex* patterns on *Stanford* parses. All four periphrastic tenses were implemented in elaborate patterns. Tense features were already implemented in Hancke (2013) in a data driven approach by pattern extraction from the *NTV* corpus, a set of 2,000 texts from the German news channel NTV[7], that was already used in Hancke, Vajjala & Meurers (2012). She decides against manually crafted patterns, assuming a better coverage of possible constructions (Hancke 2013: 45). While this is certainly true for possible variants of common tense patterns, there might be not enough data for rare tense patterns as future 2. Furthermore, the parse trees for texts from the domain of readability classification are, unlike learner texts, reliable enough for detailed patterns, so it seemed worthwhile to implement all possible patterns for readability classification. In fact, those could be implemented with high scores for proficiency, recall and F-score. Verb cluster size was measured up to V5; clusters of larger size were collected as 'V6 or more", due to their rarity. In order to account for the discussed difference in the frequency with which auxiliary, modal and main verbs take on verb complements, those instances were counted separately. The notion of *sub cluster* was introduced, referring to units of two adjacent verbs, one

[7] http://www.n-tv.de.

governing the other and ratios were collected for each sub cluster type.

3.3. Features Based on Topological Field Positions

The third feature set, called TF, is considerably smaller than the first two and is, as the following two feature sets, too, a collection of features suggested by Doreen Bryant. It contains two syllable distance measures and the ratio of non-subject prefields. Conceptually, those features were grouped together, because they focus on two elementary positions around the sentence brackets in the Topological Field model: the prefield and the middle field.

Length measures as such are common shallow features in readability classification (e.g. Crossley et al. 2010) and also included in early readability formulas (e.g. Kincaid et al. 1975). Hancke (2013), too, employs a number of features measuring sequence lengths, including syllable counts as a measure of word length. In this tradition, the average distance was measured between LSB and RSB as well as between the first argument of the verb and the main verb, if the latter was not located in C/FIN. Both these measures address the middle field, albeit the latter might also include C/FIN in case of fronted verb arguments. Instead of the number of words, the number of syllables were chosen as measure, since it gives a more accurate weight long words compared to short words. To retrieve syllable counts, the same method as in Hancke (ibid.) was used: each German syllable contains a single phonological vowel, i.e, orthographic diphthongs, adjacently repeated vowels and all voewel-e combinations were considered a single vowel (ibid.: 35). Although mostly accurate, it should be pointed out that the approach slightly overgeneralises as it fails to account for hiatus. So for example, the German adjective *reell* (realistic) is counted as a single syllable, although it is in fact two syllables *re-ell*. However, for most cases the immediate repetition of the same vowel in fact indicates elongated pronunciation, as in *Meer* (sea), so this inaccuracy was considered marginal enough to be acceptable, especially as there was no syllable counter at hand that would have solved this issue properly.

Non-subject prefields were counted, since subjects are often considered prototypical elements for German prefields and, therefore, should be easier to comprehend in this position: The preverbal position is typically to be filled with a single constituent in German declarative main clauses.[8] This can either be done via $\bar{A}$-movement

[8] However, German V1 and V3 declarative are also marginally acceptable in highly restricted circumstances. These instances are not considered here, due to their exceptional status. Please see

Feature	Formula
Avg. syll. dist. LSB – RSB	# syllables LSBs – RSBs / # LSB – RSB instances
Avg. syll. dist. arg1 – main verb	# syllables arg1 – main verbs in VC / # arg1 – main verb in VC instances
Ratio of non-subject prefields	# of non-subject prefields / # prefields

Table 3.5.: Features based on topological field positions.

or via formal movement. $\bar{A}$-movement is the movement of a constituent from the middle field to the prefield inevitably triggering a contrastive reading (Rizzi 2004). Formal movement is a term defined by Frey (2004), that describes the topicalisation of the highest element in the initial – that is non-scrambled – middle field, which leads in contrast to $\bar{A}$-movement to unmarked readings. This initial linear ordering is determined by multiple factors: animateness, shortness, givenness and definiteness are common characteristics of the first elements of the middle field (cf. Behagel 1909: 139; cf. Behagel 1930: 86). Therefore, the *Wackernagel* position, which is the highest possible position in the middle field, can only be filled with pronouns (Hoberg 1981; Reis 1986). Most often, the subject fullfills these criteria and is the highest element in the middle field (Zeman 1992), so non-subject prefields seem to be good indicators of $\bar{A}$-movement. However, sentence adverbials may also occur in the prefield without triggering a marked reading, because they cannot be topical (Frey 2005: 1). They are assumed to be base-generated in the prefield position or, in terms of the $\bar{X}$-scheme in the specifier position of the complementizer phrase (CP) (Maienborn 2001: 7). As already discussed, the difference between sentence- and verb-level modification could not be made, therefore, for now only subjects were considered for the feature. It would be desirable to enhance this feature in future work. It seems also worthwhile to investigate the performance of the non-subject prefield feature on proficiency classification, too: As Ballestracci (2010) reports, Italian speakers acquire German non-subject prefields only in the last stages of acquisition. She refers to studies on children as well as on foreign workers by Clahsen, Meisel & Pienemann (1983) and Haufe (2004).

Both syllable distance features were implemented using sentence-wise *Tregex* patterns on *Berkeley* topological field parses. The pattern for the LSB–RSB distance identifies all middle fields surrounded by filled C/FIN and VC positions. Another pattern identifies clauses, in which no finite verb is positioned in C/FIN and at least one main verb is to be found in VC of the same clause. For those instances, the first argument of the main verb is determined as well as the subject based on

Müller (2005, 2010) for further discussion of German V3 declaratives.

Feature	Formula
Ratio of *man* occurrences	# *man* occurrences / # subjects
Ratio of infinitival constructions	# infinitival constructions / # VPs
Ratio of *lassen* occurrences	# *lassen* occurrences / # VPs
Ratio of half modal clusters	# half modal clusters / # VPs
Ratio of passives	# passive constructions / # finite verbs
Ratio of quasi passives	# quasi passive constructions / # finite verbs
Ratio of participle modifiers	# past participle + # present participle / # VPs
Ratio of attributive participles	# participle I or II attributes / # NPs
Coverage deagentivation patterns	# observed deagentivation types / # possible deagentivation types

Table 3.6.: Features based on deagentivation patterns

the *Mate* dependency parses. The syllable distance is calculated between the main verb and the first in terms of linear order of those two. The ratio of non-subject prefields is calculated with the dependency parser for subject identification and the *Berkeley* topological field parses to test whether the subject is in fact in the prefield.

3.4. Features Based on Deagentivation Patterns

A common characteristic of the habitus of academic writing is the authors retreat behind the text (Kaiser 2002; Polenz 1981; Schlömer 2012). This stylistic device is commonly used to imply objectivity and general validity of the text content (Henning & Niemann 2013: 440f,Roelcke 2010: 83). Polenz (1981) coined the term *deagentivation* for this strategy of establishing legitimacy, as all those patterns are strategies to omit the agent of event predicates (ibid.: 97, Henning & Niemann 2013: 444). Henning & Niemann (Figure 1 ibid.: 447) describe several commonly identified morpho-syntactic deagentivation patterns in their investigation of non-personal style in academic writing and systematize them. They differentiate between avoidance of the person feature in the verbal category and maintenance of it. The former case may be organized verbally, by means of infinitival, participle or non-finite constructions, or nominally, by participle attributes or deverbal nominalisations. The latter compensates the personal feature in the verbal category again either verbally, by means of passive voice, half modals or *lassen* and reflexive constructions, or nominally, by means of subject shift or the usage of *man* instead of a subject properly referring to an agent.

As to be seen in Table 3.6, a subset of those patterns was implemented as features for readability classification in a feature set named Deag. It includes infinitival constructions, *lassen* and *man* occurrences, passives and half modals[9]. Nominal and verbal participle modifiers were also included using the implementations from the complex phrase feature sets. They, therefore, occur in more than one feature set. Subject shift and deverbal nominalisations are not implemented, yet, but will be approached in future work. However, additional to passive constructions, quasi passive constructions were also included, as they serve the same deagentivational function. Quasi passives are *bekommen*, *erhalten* or *kriegen* in combination with past participle (§179 Duden (Gr) 2009: 147f). Example 4a to Example 4c illustrate the similarity between German passive and quasi passive constructions.

(4) a. Man weist dir eine Stelle zu.
someone assign.3.Sg.Pres you.Dat a position *verb particle*
'Someone will assign a position to you.'

b. Dir wird eine Stelle zugewiesen.
you.Dat will.3.Sg.Pres a position assign.PP
'A position will be assigned to you.'

c. Du bekommst/erhältst/kriegst eine Stelle zugewiesen.
you.Dat obtain/receive/get.2.Sg.Pres a position assign.PP
'You will obtain/receive/get a position.'

Example 4a shows an active sentence with a *man* subject. This purely formal subject is dropped in the regular passive construction in Example 4b, with the object retaining the dative case. In Example 4c again the original agent disappears, yet, the object is assigned nominative case, serving as the syntactic subject, similar to *man*, while retaining its theta role as patient.

It should be noted that Hancke (2013: 40f) already identifies passive voice with two *Tregex* patterns. However, the earlier patterns regarded German *Vorgangspassiv* for *werden/wird/wurden*, only, ignoring future 2 passives completely. This restriction was necessary, as the patterns had to be designed requiring only minimal information about clause structure in order to allow the features to be employed for proficiency classification on learner language. Since those limitations do not apply to the current task of readability classification, the *Tregex* passive patterns were redesigned to cover the full variety of passive voice. That is, additional to German *Vorgangspassiv*, *Zustandspassiv* was included, too. For both passives, all combinations of phi features, i.e. person, number, tense, mode and voice, were covered. Also, the ratio was calculated with the number of finite verbs, while Hancke

[9] Half modals are *haben, sein, scheinen, drohen, versprechen,* if they govern an infinitive with *zu* (§101 Duden (Gr) 2009: 101).

(ibid.: 42) calculates it based on sentence as well as on clause counts. However, it seemed reasonable to normalise with the number of finite verbs, as this is the closest approximation of possible instances for passives in a text without identifying verbs that cannot form a passive. While returning high precision, recall and F-score values on the Reference corpus, those elaborate patterns are probably not suited for learner language, as they are dependant on correct inflection and precise parses. Except for the number of infinitival constructions, which was identified using the POS information provided by the *Mate* parser, all other counts were also collected with rather straight forward *Tregex* patterns.

3.5. Features Based on Conditional Clauses

The last feature set, referred to as Cond, addresses the mediation and integration of German conditional clauses. Graesser et al. (2004: 198) employ conditionality markers *wenn* and *dann* to measure logical and analytical complexity of a text. However, conditional clauses may differ in how explicit the protasis, i.e. the clause expressing the condition, is marked. Conditional marking may be introduced with the conditional subjunction *wenn* (if) as in Example 5. However, it may as well remain unmarked, leading to an implicit protasis in form of a V1 clause, see Example 6. Such unmediated V1 clauses are typically conditional clauses, although in rare cases other types of adverbials may also be expressed with this construction, see Freywald (2013) for discussion. However, due to their marginality they were ignored for this feature set.

(5) Wenn du gerne Kuchen isst, [apodosis]
if you gladly cake eat.2.Sg.Pres [apodosis]
If you like to eat cake, [apodosis]

(6) Isst du gerne Kuchen, [apodosis]
eat.2.Sg.Pres you gladly cake [apodosis]
'If you like to eat cake, [apodosis]'

Furthermore, there are three different levels of integration of the protasis to the apodosis, i.e. the consequence bearing main clause: Complete incorporation is indicated by a V1 apodosis in Germanic languages (König & van der Auwera 1988: 102). This is illustrated in Examle 7a, where protasis and apodosis form a V2 sentence with the entire protasis in the prefield. These constructions are called *integrative Spitzenstellung* (integrative conditionals).

(7) a. [protasis], musst du auch backen können.
[protasis], must.2.Sg.Pres you also bake.Inf can.Inf

[protasis], you also have to be able to bake.

b. [protasis], dann musst du auch backen können.
[protasis], then must.2.Sg.Pres you also bake.Inf can.Inf

[protasis], then you also have to be able to bake.

In Example 7b, the apodosis is introduced by the resumptive element *dann*. The protasis is located in the pre-prefield. Accordingly, those constructions are referred to as *resumptive Spitzenstellung* (resumptive conditionals). Resumptive *wenn*-initial conditional clauses, in the following referred to as *wenn-dann conditionals*, are the most explicitly marked German conditionals. In contrast, integrative V1-initial conditionals, in the following referred to as *V1-V1 conditionals*, trigger implicit inferences. So by counting explicit signals of conditionals only, i.e. *wenn* and *dann* instances, not all conditional clauses are included. In fact, those conditionals requiring the reader to establish coherence of a text without unambiguous, explicit cohesion markers[10] are omitted. This is especially problematic, since Axel (2002: 9) reports findings from a study performed on several corpora of contemporary German texts[11], in which 85% of the conditionals were integrated and only 14% resumptive. For the sake of completeness, it should be mentioned, that albeit these numbers, König & van der Auwera (1988: 117) as well as Freywald (2013) assume, that in modern German the variant with *dann* is preferred. Regardless of the detailed preference ranking of resumptive and integrated conditionals, it is clearly desirable to cover the latter as well in reading classification. Therefore, Cond was designed to model conditionals with and without *dann*, each with both types of protases, that is with and without *wenn*.

[10] Based on the distinction made by Graesser et al. (2004) cohesion is understood as "an objective property of the explicit language and text" created by "explicit features, words, phrases, or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units (e.g., topics and themes)", while coherence refers to "a characteristic of the reader's mental representation of the text content" (ibid.: 193).

[11] Those corpora were:

1. the first 100 instances of adverbial clauses in the left sentence periphery in the weekly news magazine *Der Spiegel* (`http://www.spiegel.de/`), No. 9/2000, p. 1–151,
2. the novel *Von allem Anfang an* by C. Heins (Berlin 1977),
3. the first 50 instances of adverbial clauses in the left sentence periphery in the German daily newspaper *Süddeutsche Zeitung*, No. 49/2000, p. 1–4, 13–18,
4. the first 50 instances of adverbial clauses in the left sentence periphery in the text collection *Celebration* by R. Goetz (Frankfurt/Main 1999).

Features	Formula
Ratio *wenn-dann* conditionals	# *wenn-dann* conditionals / # conditionals clauses
Ratio *wenn*-V1 conditionals	# *wenn*-V1 coditionals / # conditional clauses
Ratio of V1-*dann* conditionals	# V1-*dann* conditionals / # conditional clauses
Ratio V1-V1 conditionals	# V1-V1 conditionals / # conditional clauses
Coverage conditional clause types	# observed conditional clause types / # possible conditional clause types

Table 3.7.: Features based on conditional clauses

For sake of completeness, a third type of conditional clauses should be mentioned. It is shown in Example 8.

(8) Wenn du mitkommen willst, ich habe nichts
if you come-along.Inf want.2.Sg.Pres, I have.1.Sg.Pres nothing
dagegen. ((28) König & van der Auwera 1988: 115)
against-it
'If you want to come along, I don't mind.'

In this sentence, the apodosis is a complete clause by itself, that is, the protasis is not integrated at all. These conditionals are typically referred to as *nicht-integrative Spitzenstellung* (non-integrated conditionals), in which protasis and apodosis are the least conjoined. König & van der Auwera (ibid.) also refer to them as non-canonic conditionals, because they are least common and only licensed in highly restricted contexts (Lötscher 2006: 148). They also occur only with 2% in Axel (2002: 9)'s corpus study. Interestingly, in historic German non-integrative conditionals were not restricted, but in free variance with resumptive conditionals (Lötscher 2006: 349). König & van der Auwera (1988: 108) suggest that the turn from non-integrated to integrated protases is due to the developement of German to a stricter V2 language. However, since in modern German non-integrated conditionals are only marginal constructions, they were not considered in the feature set. For further theoretical discussion please see Axel (2002), König & van der Auwera (1988), and Lötscher (2006).

Table 3.7 lists the features and formulas implemented in this feature set. As already mentioned, resumptive and integrated conditionals were counted with and without *wenn* in the protases. This was realised with four *Tregex* patterns on topological field parses, probing for a) occurrences of *wenn* in C/FIN or b) *dann* in the prefield of a clause with an adjacent clause in the same sentence or c) for two adjacent V1 clauses in a declarative sentence. The latter was considered sufficient, notwithstanding the briefly mentioned exceptions, since most V1-V1 declaratives are in fact conditionals. However, the actual performance of the patterns could not be measured adequately in the Reference corpus due to the low counts for conditionals, which only contained five *wenn-V1* conditionals and two *V1-V1* conditionals. None of the other conditional clause types occurred. While precision, recall and F-score for the former were satisfying, the latter only reached an F-score of 0.667. However, the parse for the conditional clause in question was highly skewed and, therefore, not considered representative, which is why feature and pattern were retained albeit the low performance on the Reference corpus. Also, the coverage of conditonal types in a text was measured to access variance.

4. Classification Experiments

4.1. Set-up

Two classification experiments were conducted to test the accuracy of the new features in a readability classifier: one classifying school types and one classifying grade levels. From a theoretical point of view a measurable difference in readability could be expected for both dimensions: Language complexity should ascend with increasing grade levels and thereby adjust to the student's increasing language proficiency. Regarding school types language complexity might also differ due to the different qualification foci set by *Gymnasium* and *Hauptschule*. Special attention was paid to differences between publishers, due to the significant differences among those reported for the Reading Demands corpus in Vajjala (2015), chapter 8.

The experiments were performed on the Full corpus and the Publisher A to D corpora introduced in section 2.3 using the Waikato Environment for Knowledge Analysis (WEKA)[1] software. WEKA is a Machine Learning workbench providing several visualisation tools and Machine Learning algorithms for data analysis (Witten, Frank & Hall 2011). In the course of the experimental set-up several Machine Learning algorithms implemented in WEKA were tested using 10 folds cross-validation, including Naive Bayes, Linear as well as Multinomial Logistic Regression and Support Vector Machines (SVM) from the libsvm package (Chang & Lin 2013). Ultimately, the Sequential Minimal Optimization (SMO) algorithm by John Platt was employed. It returned the best results and was also used in earlier experiments with the original classifier, e.g. in Galasso (2014) and Hancke, Vajjala & Meurers (2012). Afterwards, the separate features were ranked with the information gain algorithm from WEKA to retrieve more detailed information on their individual merit. Again ten folds cross-validation was used.

Three baselines were established to evaluate the new features: First, a random baseline (Rand) to show whether the features allow for more accurate classification than by assigning classes by chance. For binary school-wise classification this is

[1] `http://www.cs.waikato.ac.nz/~ml/index.html`.

Feature set	# Features	Grade-Wise		School-Wise	
		Full corpus	Full corpus + Publ	Full corpus	Full corpus + Publ
Syn	47	44.83%	46.25%	64.34%	69.08%
Lex	46	48.56%	49.33%	68.25%	71.46%
Morph	42	47.01%	49.74%	61.74%	68.04%
Tran	16	37.18%	40.44%	54.24%	62.44%
Ref	8	37.46%	40.44%	55.00%	62.50%
PrepDet	7	37.46%	40.44%	52.61%	62.54%
Descr	3	37.43%	40.54%	64.30%	66.17%
Conn	6	37.46%	40.44%	52.61%	62.54%
Rand	0	33.33%	33.33%	50.00%	50.00%
Orig 2013	135	52.23%	53.41%	70.22%	73.78%
Orig 2014	175	53.17%	53.23%	71.60%	76.10%

Table 4.1.: Performance of the original classfier on grade-wise and school-wise classification.

50.00% and for ternary grade-wise classification it is 33.33%. The second baseline is the performance of the original classifier by Hancke (2013). As this classifier has already been enhanced by Galasso (2014) a third baseline was established based on this enhancement. While the second baseline indicates the improvement caused by the new linguistically insightful features, the third baseline is used to indicate whether the features remain beneficial for a classifier that already employs features of deeper linguistic insight.

4.2. Feature Set Performances

4.2.1. Performance of the Baseline Classifiers

For each of the classification experiments the results for the original classifiers from 2013 (Orig 2013) and 2014 (Orig 2014) were replicated. All three baselines are displayed as feature sets in Table 4.1 together with an overview over the performance of the different feature sets in both original classifiers. In each experiment, the addition of the publisher feature increases the classification accuracy further. The feature sets from the original classifier by Hancke (2013) perform with high accuracy in both, grade-wise and school-wise classification, with Lex as the most predicting feature set. While Morph performs good average, Syn is considerably more distinctive for school-wise classification, being distant third for grade-wise classification. However, all three feature set outperform the additional features implemented by Galasso (2014), except for the superficial Descr feature. That was to be expected, since those feature sets were designed as complements to the original feature sets. As such they improve the classifier by 0.94% for the grade-wise and 2.32% for the school-wise classification. Interestingly, all linguistically informed feature sets by Galasso (2014) return approximately equal accuracies. Also, the publisher feature increases accuracy for school-wise classification considerably. However, for grade-wise classification Orig 2014 returns lower accuracies than Orig 2013 given the publisher feature, although the feature overall increases the accuracy of the separate feature sets. It seems worthwhile to investigate these unexpected results further by performing experiments on the separate Publisher corpora. However, this would be beyond the scope of this thesis, which focusses on the newly implemented feature sets.

4.2.2. Performance of the Enhanced Classifiers

All three baselines are repeated in Table 4.2 for grade-wise classification and in Table 4.3 for school-wise classification. For each classification experiment, two configurations of the enhanced classifier are compared to their respective baseline: Enha 2013 enhances Orig 2013 and Enha 2014 enhances Orig 2014. Also, each enhanced classifier was configured once by feature reduction. The features performing worst in the respective classification experiment in terms of information gain were deleted from the enhanced classifier, resulting in the classifiers Enha 2013 / 2014 - Worst 6 SW / Worst 7 GW). The underlying feature ranking is discussed in section 4.2.3 in Table 4.4 and Table 4.3. The number of features to be removed was calibrated such that it increased accuracy as much as possible. For each feature set in the classification experiments, the number of features contained by it is reported. The accuracy in both experiments is reported for each Publisher corpus and for the Full corpus, the latter once with and once without the additional publisher feature. The highest accuracy for each corpus in each classification experiment configuration is marked with bold font. If the highest accuracy was achieved with multiple feature sets, the set containing the least features was marked as best performing feature set.

Grade-Wise Classification On first sight grade-wise classification in Table 4.2 does not benefit significantly from the new features on the Full corpus. Although Enha 2013 and Enha 2014 beat the random baseline, accuracy actually drops for both versions of the enhanced classifier in comparison to Orig 2013 and Orig 2014. Enha 2014 performs even slightly worse than Orig 2013. Yet, in combination with the publisher feature the results reverse and the new features return up to 0.52% and 0.42% higher accuracy values than the original classifiers. Furthermore, Enha

Feature set	# Features	Publisher A	Publisher B	Publisher C	Publisher D	Full corpus	Full corpus + Publ
Baseline I: Rand	0	33.33%	33.33%	33.33%	33.33%	33.33%	33.33%
Baseline II: Orig 2013	135	53.41%	**55.56%**	**62.79%**	**61.56%**	**52.23%**	53.41%
Enha 2013	197	53.51%	54.91%	61.03%	60.36%	51.94%	53.75%
Enha 2013 - Worst 7 GW	172	**53.61%**	52.98%	61.91%	61.56%	52.13%	**53.93%**
Baseline III: Orig 2014	175	**54.87%**	56.20%	**60.70%**	**63.94%**	**53.17%**	53.23%
Enha 2014	219	53.12%	**57.00%**	60.59%	62.46%	52.20%	**53.65%**
Enha 2014 - Worst 7 GW	212	53.70%	55.56%	61.03%	63.94%	51.71%	53.61%

Table 4.2.: Performance baselines and overall performance for grade-wise classification.

2013 returns 0.10% higher accuracy than Enha 2014. This unexpected effect of the publisher feature can be explained by taking a closer look at the performance differences among the Publisher corpora. Enha 2013 increases accuracy only for the Publisher A corpus by up to 0.20%. For all other Publisher corpora, the Orig 2013 outperforms both version of Enha 2013. The situation for Enha 2014 is comparable: Accuracy increases by 0.80% on the Publisher B corpus, yet it decreases on all other Publisher corpora. This suggests that for the Publisher C and D corpus the linguistic properties described by the new features are not used to differentiate readability levels between grades. However, at least for the Publisher C corpus this seems to be part of a more general lack of more complex linguistic differences between the different grades, since accuracy also drops from Orig 2013 to Orig 2014. As for the Publisher A corpus: it seems to be the case that the difference between grades manifests also in more complex linguistic features, yet the features from Galasso (2014) capture these differences more accurately, while in the Publisher B corpus, the combination of both feature sets seems to have some beneficial cumulative effect leading to the higher accuracy. This would also explain the considerable drop of accuracy between Enha 2014 and Enha 2014 - Worst 7 GW. The fact that the new features increase accuracy for only one Publisher corpus each explains the results for the Full corpus without the publisher feature. However, given the information on the publisher, the beneficial effect of Enha 2013 - Worst 7 GW in the largest of the four Publisher corpora shows on the Full corpus, too. The same holds for Enha 2014; where the Publisher B corpus is smaller, yet the increase in accuracy also is considerably higher, leading to similar results on the Full corpus.

School-Wise Classification Accuracy for school-wise classification in Table 4.3 is in general higher compared to grade-wise classification. Obviously, this most prominent difference between the two experiments is due to the difference between

Feature set	# Features	Publisher A	Publisher B	Publisher C	Publisher D	Full corpus	Full corpus + Publ
Baseline I: Rand	0	50.00%	50.00%	50.00%	–	50.00%	50.00%
Baseline II: Orig 2013	135	65.11%	**80.68**%	77.39%	–	70.22%	73.78%
Enha 2013	197	**65.50**%	79.55%	76.62%	–	**70.91**%	**75.06**%
Enha 2013 - Worst 6 SW	172	65.01%	80.52%	**77.72**%	–	70.49%	74.82%
Baseline III: Orig 2014	175	**66.47**%	**80.19**%	**81.12**%	–	71.60%	76.10%
Enha 2014	219	65.98%	79.23%	80.24%	–	**71.84**%	76.48%
Enha 2014 - Worst 6 SW	213	65.69%	79.55%	80.35%	–	71.67%	**76.86**%

Table 4.3.: Performance baselines and overall performance for school-wise classification.

binary and ternary classification and is, therefore, not relevant for the discussion. Unlike with the grade-wise classification experiment, the new features lead to improved accuracy for school-wise classification on the Full corpus without as well as with the additional publisher feature: Enha 2013 leads to 0.69% higher accuracy without and 1.28% with the publisher feature and Enha 2014 to 0.24% higher accuracy without and 0.76% with it. However, the publisher feature improves accuracy by up to 5.02% for the Full corpus. This extreme effect of the publisher feature might be explained by the difference in accuracy between the Publisher A corpus and the other Publisher corpora: While classification on the Publisher B and C corpora performs nearly comparably ranging from 77.71% to 81.12%, performance drops considerably on the Publisher A corpus, ranging from 65.50% to 66.47%. The comparably low accuracy values for school-wise classification on the Publisher A corpus indicate that the linguistic difference between *Gymnasium* and *Hauptschule* is less prominent in this corpus. Interestingly, this does not imply that the more linguistically informed features do not improve classification: Enha 2013 improves accuracy the most for the Publisher A and C corpora with 0.39% and 0.33%. Also, Orig 2014 performs better on the Publisher A corpus than Orig 2013, which also indicates that more linguistically motivated features have a beneficial effect on the classification. As in the previous experiment on the Publisher A corpus, the new features fail to improve accuracy on the corpus in comparison to Orig 2014. The same holds for the Publisher C corpus. Yet, accuracy does not drop from Orig2013 to Orig2014, which suggests that the different school types, unlike the grade levels, in general can be distinguished linguistically in the Publisher C corpus. For the Publisher B corpus, in neither experiment an improve in accuracy could be induced by the new features. Also, accuracy dropped from Orig 2013 to Orig 2014, which suggests that the difference between the two school types is not encoded by more linguistically motivated features in general in this corpus. Overall, the results

Feature set	# Features	Publisher A	Publisher B	Publisher C	Publisher D	Full corpus	Full corpus + Publ
CompNP	8	**45.42**%	42.35%	**49.62**%	48.94%	**43.69**%	**45.59**%
CompVP	21	44.93%	**47.83**%	43.36%	**50.15**%	37.67%	40.09%
TF	3	44.93%	43.00%	43.14%	49.25%	**40.92**%	**41.61**%
Deag	9	**45.03**%	43.96%	**44.90**%	**50.45**%	37.36%	40.40%
Cond	5	44.54%	**45.41**%	44.13%	49.55%	37.84%	40.47%
Best 2 Features	–	46.49%	46.70%	51.15%	47.45%	43.69%	45.59%
CompNP + CompVP	30	46.20%	48.63%	48.08%	48.95%	44.10%	45.90%
CompNP + CompVP + Deag	36	**48.15**%	49.76%	51.15%	**53.75**%	43.17%	45.90%
All	44	46.59%	**51.05**%	51.04%	51.95%	45.69%	**47.60**%
All - Worst 7 GW	37	46.49%	49.11%	**52.36**%	50.75%	**45.83**%	47.11%

Table 4.4.: Performance overview for feature sets and feature set combinations on grade-wise classification.

on the Full corpus mirror the results on the Publisher corpora for Enha 2013. The results on the Publisher B corpus do not counterbalance the results from the other two Publisher corpora, since it is the smallest of the three. The improved accuracy on the Full corpus for Enha 2014 is less straight forward to explain. It seems as if the features from Galasso (2014) fit the data in the separate corpora better, yet on the Full corpus the linguistic properties on the texts are cumulated such that the new features can add beneficial information to the classification task.

4.2.3. Performance of the Separate New Feature Sets

Additional to the overall performance of the new features, the performance of the separate feature sets introduced in chapter 3 is of special interest. Their accuracy on the separate Publisher corpora and the Full corpus is displayed in Table 4.4 and Table 4.5, which are designed similar to the previous tables on the overall performance. However, in the single feature sets the two best accuracy values are marked with bold font, to increase the readability of the feature set combination Best 2 Features. Both tables show the performance of several such combinations of feature sets additional to the single feature sets. The two sets All and All - Worst 7 GW / 6 SW refer to the feature sets that were used to build the enhanced classifiers reported on in section 4.2.2.

Grade-Wise Classification Table 4.4 shows that the feature set describing complex NPs is most predictive for grade-wise classification on the Full corpus with and without the additional publisher feature. Distant second is the TF feature set, followed by the other three feature sets with approximately the same accuracy. Together, both feature sets account for 43.69%, respectively 45.59% of the accuracy

Feature set	# Features	Publisher A	Publisher B	Publisher C	Publisher D	Full corpus	Full corpus + Publ
CompNP	8	59.65%	66.18%	**69.05**%	–	**59.63**%	**66.24**%
CompVP	21	59.75%	66.67%	63.45%	–	57.77%	63.09%
TF	3	**59.75**%	**66.67**%	63.45%	–	57.90%	62.88%
Deag	9	59.75%	**66.99**%	**63.78**%	–	**59.91**%	**63.54**%
Cond	5	**59.75**%	60.55%	63.45%	–	52.37%	62.47%
Best 2 Features	–	59.75%	69.73%	**71.79**%	–	59.53%	65.72%
CompNP + CompVP	30	59.36%	71.01%	68.39%	–	62.02%	68.49%
CompNP + CompVP + Deag	36	60.43%	72.46%	70.47%	–	63.92%	68.66%
All	44	**61.01**%	**74.40**%	71.57%	–	64.82%	**70.32**%
All - Worst 6 SW	38	60.14%	73.27%	71.46%	–	**65.13**%	70.05%

Table 4.5.: Performance overview for feature sets and feature set combinations on school-wise classification.

on the Full corpus, as Best 2 Features shows. Also, CompNP and TF are the only feature sets on the Full corpus that do not classify most of the data as 5^{th} and 6^{th} grade as the other feature sets do, but return a rather well-formed confusion matrix.Yet, a tendency to classify texts as 5^{th} and 6^{th} grade texts remains. The high performance of CompNP seems to be due to its expressiveness for the two largest Publisher corpora. On the smaller Publisher B and C corpora, however, it actually returns the lowest accuracy. This suggests, that the usefulness of the CompNP features is rather corpus dependant than a general indicator of grade levels. Since CompVP ranges among the two best feature sets for the other two corpora, CompNP and CompVP were collected in a joint feature set CompNP + CompVP, which lead to an average performance on all corpora. Although TF is the second most predicting feature on the Full corpus, it is on neither Publisher corpus. Instead, Deag is among the two most predicting feature sets on three of the four publisher corpora. Accordingly, combining CompNP + CompVP with Deag leads to the highest accuracy for the Publisher A and D corpora. On the other two corpora, some version of the full feature set returned the highest accuracy, just as with the Full corpus.

School-Wise Classification The results for school-wise classification are similar to those for grade-wise classification. Again, CompNP is one of the most predictable features on the Full corpus and on the Publisher C corpus, while Deag performs second best. As in previous experiments, the additional publisher feature increases accuracy considerably. The high accuracy for CompNP on the Full corpus seems to be motivated by its exceptional high performance on the Publisher C corpus, supporting the analysis from the grade-wise classification. While the feature performs worst on the other two Publisher corpora, with 69.05% it returns results comparable to the feature set combinations on the Publisher C corpus. Distant second best with 5.27% less accuracy is Deag. However, except for Cond all feature sets perform comparable to Deag on this corpus, which is a difference to the results on the Full corpus. Accuracy values on the other two Publisher corpora are similarly close. On the Publisher D corpus, TF and Deag perform best. Yet, except for Cond, which returns remarkably low accuracies on this corpus, all accuracies are between 66.55% and 66.99%. As for the Publisher A corpus, neither feature returns any remarkable accuracy. Instead, all perform comparably low with not quite 60% accuracy. The marginal differences between most feature sets on the Publisher corpora explains how neither of the feature set combinations performs better than All. Only for the Publisher C corpus Best 2 Features returns slightly higher results than All.

4.3. Feature Rankings

After discussing the overall performance of the new features and of the separate feature sets all features were ranked separately in terms of information gain, in order to get a deeper insight into the results. This was necessary, since it is not immediately transparent what it means linguistically, when CompNP proofs to be the most predicting feature. This is illustrated in Table 4.6, where three CompNP features are ranked among the best and three among the worst five features in terms of information gain. Therefore, first the most and least predicting features from all new feature sets for both classification tasks are briefly presented, followed by an in depth discussion of the feature set internal rankings.

Table 4.6 and Table 4.7 display the feature rankings over all new feature sets for grade-wise and school-wise classification. In each table, the five highest and the seven, respectively six lowest ranked features are shown together with their corresponding feature sets, their average merit and the standard deviation σ. To give some context to the values displayed in the table, the average merit for the full Enha 2014 classifier should be outlined first: Average merit for all features in Enha 2014 ranges from 0.0 to 0.75 for grade-wise classification. The median of the average merits of all features above 0.0 is 0.017. For school-wise classification average merit ranges from 0.0 to 0.053 with a median of 0.013. Interestingly, this pattern of higher average merits for grade-wise classification on Enha 2014 is inverted for the new feature sets. For them, higher merits are achieved for school-wise classification throughout all rankings except for the conditional clause features.

Rank	Feature Set	Average Merit (±σ)	Feature
1.	TF	0.023 (± 0.002)	average syllable distance LSB/RSB
2.	CompNP	0.020 (± 0.001)	ratio postnominal modifiers
3.	CompNP	0.019 (± 0.001)	ratio possessive modifiers
4.	TF	0.017 (± 0.002)	average syllable distance arg1 to main verb
5.	CompNP	0.011 (± 0.001)	ratio prenominal modifiers
...	...	...	...
37.	CompVP	0.000 (± 0.000)	coverage tense
38.	CompNP	0.000 (± 0.000)	coverage modifier types
39.	CompNP/Deag	0.000 (± 0.000)	ratio attributive participles
40.	CompVP	0.000 (± 0.000)	coverage periphrastic tense
41.	CompNP	0.000 (± 0.000)	ratio determiners
42.	Cond	0.000 (± 0.000)	ratio V1-V1 conditionals
43.	CompNP	0.000 (± 0.000)	ratio comparative modifiers

Table 4.6.: Best and worst features for grade-wise classification from all feature sets in terms of information gain for all publishers.

Rank	Feature Set	Average Merit (±σ)	Feature
1.	Deag	0.029 (± 0.002)	coverage deagentivation patterns
2.	Comp NP/Deag	0.026 (± 0.003)	ratio attributive participles
3.	TF	0.021 (± 0.002)	average syllable distance LSB/RSB
4.	Comp NP	0.018 (± 0.001)	ratio clausal modifiers
5.	Comp NP	0.018 (± 0.003)	ratio prenonminal modifiers
...	...	...	...
38.	TF	0.000 (± 0.000)	ratio non subject prefields
39.	CompVP	0.000 (± 0.000)	ratio future 2
40.	Deag	0.000 (± 0.000)	ratio half modal clusters
41.	Deag	0.000 (± 0.000)	ratio *lassen* occurrences
42.	CompVP	0.000 (± 0.000)	ratio past perfect
43.	Cond	0.000 (± 0.000)	coverage conditional types

Table 4.7.: Best and worst features for school-wise classification from all feature sets in terms of information gain for all publishers.

On both classification tasks, some CompNP features score: For grade-wise classification, these are pre- and postnominal as well as possessive modifiers. However, coverage of modifier types, ratio of determiners, ratio of attributive participles and ratio of comparative modifiers are among the worst seven features. Interestingly, additional to prenominal modifiers participle modifiers are among the best five features for school-wise classification, as well as clausal modifiers, indicating intriguing differences between grade-wise and school-wise classification. With the average syllable distance between LSB and RSB as well as between first argument and main verb TF, too, ranks high on grade-wise classification. While the length of the middle field in terms of syllables proofs to be highly predictive for both classification tasks, the syllable distance between first argument and main verb only ranks among the top five features for grade-wise classification. The ranking, thereby, matches the results from the feature set classifications in Table 4.4 and Table 4.5, where TF worked better on grade- than on school-wise classification. Instead of TF, Deag features rank higher for school-wise classification, which is again consistent with the results in Table 4.5. Whether participle modifiers score high because of their deagentivational function remains undecided. If so the absence of participle verb modifiers would have to be explained. Also, as the low ranking for half modal clusters and occurrences of *lassen* indicate, it has to be discussed further which Deag features are actually beneficial. These issues will be addressed in the respective in depth discussions of the single feature set rankings. Considering the lowest ranks on both tables, tense or variance measuring features in general seem to be rather unpredictive. However, the data shown so far does not suffice to discuss this further, which is why all feature sets were also ranked separately and will now be discussed in detail.

4.3.1. Complex Noun Phrase

Table 4.8 shows the ranking for CompNP features in grade-wise and school-wise classification. As Hancke (2013) already implemented two features measuring general complexity of NPs, those two features from the Syn feature set were also listed in the table for comparison, yet excluded from the actual ranking. Average length of NPs in terms of words is for both classification tasks the most predictive feature. This coincides with the high scores for all length measures in general. For school-wise classification, also the average number of modifiers in NPs scores higher than the ratios for the single modifier types. The difference between modifier types seems, therefore, to be less crucial for school-wise classification, than the

Rank	Grade-Wise		School-Wise	
	Average Merit ($\pm\sigma$)	Feature	Average Merit ($\pm\sigma$)	Feature
–	0.022 (± 0.002)	average NP length in words	0.034 (± 0.004)	average NP length in words
–	–	–	0.027 (± 0.003)	average number of modifiers per NP
1.	0.020 (± 0.001)	ratio prenominal modifiers	0.026 (± 0.003)	ratio attributive participles
2.	0.019 (± 0.001)	ratio possessive modifiers	0.018 (± 0.001)	ratio clausal modifiers
–	0.017 (± 0.004)	average number of modifiers per NP	–	–
3.	0.011 (± 0.001)	ratio prenomninal modifiers	0.018 (± 0.003)	ratio prenominal modifiers
4.	0.006 (± 0.003)	ratio clausal modifiers	0.013 (± 0.002)	ratio postnominal modifiers
5.	0.000 (± 0.000)	ratio attributive participles	0.012 (± 0.002)	ratio possessive modifiers
6.	0.000 (± 0.000)	ratio comparative modifiers	0.011 (± 0.001)	ratio determiners
7.	0.000 (± 0.000)	ratio determiners	0.007 (± 0.006)	ratio comparative modifiers
8.	0.000 (± 0.000)	coverage modifier types	0.010 (± 0.001)	coverage modifiers types

Table 4.8.: Ranking of complex NP features for grade-wise and school-wise classification in terms of information gain

amount of modification in general. This is consistent with the comparably equally well performance of the separate CompNP features for this classification task: While for school-wise classification all CompNP features are beneficial to some extent, for grade-wise classification only post- and prenominal as well as possessive and clausal modifiers return an average merit above 0.0. Furthermore, there is a significant drop from possessive to prenominal and from prenominal to clausal modifiers. In accordance with this more diverse picture for grade-wise classification, the number of modifiers per NP is less predictive than the ratio of prenominal and possessive modifiers. As for the publisher corpra: postnominal modifiers are consistently highly predictive on the Publisher corpora A to C, see Appendix C, Tables C.1 to C.6, pages 71f, while the informativeness of the other CompNP features varies dependant on the publisher. For example, on the Publisher C corpus, the ratio of determiners is in fact highly informative for school-wise classification as is for the Publisher B corpus the ratio of possessive modifier. These differences show that it is informative to model the domain of the complex NP in greater detail. This is especially true for grade-wise classification, but obviously that is not to say the CompNP features were not also beneficial for school-wise classification. In fact, the values for average merit are generally increased for school-wise classification. Finally, it seems worthwhile to mention that notwithstanding these differences, in terms of ranks both classification experiments return similar results: The four best features for grade-wise classification are also among the five best school-wise classification. The same holds for the three worst features. The only significant difference in ranking is the ranking for attributive participles. This difference was already briefly addressed in the previous section and appears even more prominently in this table: while the feature does not contribute to grade-wise

Rank	Grade-Wise		School-Wise	
	Average Merit (±σ)	Feature	Average Merit (±σ)	Feature
–	0.007 (± 0..03)	average VP length in words	0.012 (± 0.001)	average number of VP modifiers
–	–	–	0.013 (± 0.002)	average VP length in words
1.	0.005 (± 0.004)	ratio adjectival/adverbal modifiers	0.011 (± 0.002)	variance cluster size
2.	0.006 (± 0.003)	ratio future 1	0.010 (± 0.001)	average cluster size
3.	0.000 (± 0.000)	average cluster size	0.009 (± 0.003)	ratio periphrastic tense
4.	0.000 (± 0.000)	prepostional modifiers	0.006 (± 0.001)	coverage cluster types
5.	0.000 (± 0.000)	verb particles	0.005 (± 0.002)	ratio prepositional modifiers
6.	0.000 (± 0.000)	ratio phi feature sub clusters	0.008 (± 0.006)	ratio present perfect
7.	0.000 (± 0.000)	ratio main verb sub clusters	0.005 (± 0.002)	ratio simple past
8.	0.000 (± 0.000)	ratio modal verb sub clusters	0.002 (± 0.003)	ratio adjectival/adverbial modifiers
9.	0.000 (± 0.000)	variance cluster size	0.004 (± 0.002)	ratio future 1
10.	0.000 (± 0.000)	ratio participle modifiers	0.000 (± 0.000)	ratio modal verb sub clusters
11.	0.000 (± 0.000)	coverage tense	0.000 (± 0.000)	ratio verb particles
12.	0.000 (± 0.000)	coverage periphrastic tense	0.002 (± 0.003)	ratio participle modifiers
13.	0.000 (± 0.000)	ratio simple present	0.000 (± 0.000)	ratio main verb sub clusters
14.	0.000 (± 0.000)	ratio simple past	0.000 (± 0.000)	ratio phi feature sub clusters
–	0.001 (± 0.003)	average number of modifiers per VP	–	–
15.	0.000 (± 0.000)	coverage sub cluster types	0.000 (± 0.000)	coverage periphrastic tense
16.	0.000 (± 0.000)	ratio future 2	0.000 (± 0.000)	ratio simple present
17.	0.000 (± 0.000)	ratio past perfect	0.000 (± 0.000)	coverage tense
18.	0.000 (± 0.000)	coverage modifier types	0.000 (± 0.000)	ratio future 2
19.	0.000 (± 0.000)	ratio present perfect	0.000 (± 0.000)	ratio past perfect
20.	0.000 (± 0.000)	ratio periphrastic tense	0.000 (± 0.000)	coverage modifier types
21.	0.000 (± 0.000)	coverage cluster size	0.000 (± 0.000)	coverage cluster size

Table 4.9.: Best and worst five complex VP features for grade-wise and school-wise classification in terms of information gain

classification, it is by far the most informative feature for school-wise classification.

4.3.2. Complex Verb Phrase

Table 4.9 shows the ranking for CompVP features. Overall, the average merit is lower than for CompNP, which is straight forward since no CompNP feature ranked among the best five features and CompVP performed average in terms of classification accuracy in the feature set performance comparison in Table 4.4 and Table 4.5. As for CompNP, two features from Hancke (2013) were included in the ranking for comparison. The result is similar to the one for CompNP: the average length of VPs scores highest. While for school-wise classification the average number of VP modifiers returns a slightly higher average merit than the CompVP features, it ranks lower than several other features for grade-wise classification. Also, again all features perform better on school-wise than on grade-wise classification: for the latter only the ratio of adjectival / adverbial modifiers and the ratio of future 1 are informative at all, albeit with comparably low merits. For school-wise classification the best nine features are consistently informative. Especially descriptive information on verb clusters is beneficial, that is average cluster size as well as coverage and variance of cluster sizes. In fact, for the Publisher D

Rank	Grade-Wise		School-Wise	
	Average Merit (±σ)	Feature	Average Merit (±σ)	Feature
–	0.034 (± 0.004)	average sentence length	0.053 (± 0.003)	average sentence length
–	0.033 (± 0.004)	average T-unit length	0.044 (± 0.002)	average T-unit length
–	–	–	0.025 (± 0.002)	average clause length
1.	0.023 (± 0.002)	avg. syllable dist. LSB/RSB	0.021 (± 0.002)	avg. syllable dist. LSB/RSB
–	0.022 (± 0.003)	average clause length	–	–
2.	0.017 (± 0.002)	avg. syllable dist. arg1 to main verb	0.013 (± 0.001)	avg. syllable dist. arg1 to main verb
3.	0.000 (± 0.000)	ratio non subject prefields	0.000 (± 0.000)	ratio non subject prefields

Table 4.10.: Performance of Topological Field position features for grade-wise and school-wise classification in terms of information gain

corpus the coverage of cluster size is the third most informative feature for grade-wise classification, with an average merit of 0.044, see Appendix C, Table C.7, page 73. Also the ratio of periphrastic tenses is informative, which is correlated with higher average cluster sizes. Less important is the coverage of cluster types and the specific ratios for phi feature and modal and main verb sub clusters. Furthermore, all modifying enhancements of VPs are relevant, that is prepositional, adjectival/adverbial and participle modifiers. Yet, the average merits are far lower than for the noun modifiers in CompNP. As for tense, present perfect and simple past are ranked high as well as future 1, while simple present, future 2 and past perfect are ranked rather low. While the former is probably too common to help differentiating the different school types, the latter two might be too rare. For example, only 8 instances of future 2 were found in the entire Full corpus. The coverage of different tenses within a text proofed irrelevant for the classification task. Finally, the ratio of separated verb particles is not beneficial. However, this might be due to the highly selective acquisition of instances, as verb particles attached to their verbs were not counted. It is left for future work to determine whether the ratio complex verbs in general returns more promising results.

4.3.3. Topological Field Positions

Table 4.10 shows the ranking of the TF features. Additionally, three features measuring the length of sequences from Syn and Trad are displayed. They serve as contextualising information for the syllable distance features from TF, which perform among the best five features in both classification experiments. All sequence length features were listed but excluded from the actual ranking. The comparison shows, that average sentence length and T-uni length by far surpass the syllable distance measures for both classification tasks. This indicates that the syllable distance measures at least partially perform so well because they measure longer sequences in general. However, those two features seem to be less correlated with the syllable distance features from TF than the average clause length. This can be seen by looking at the feature rankings from the Publisher corpora A to C (Appendix C, Tables C.1 to C.6, p. 71f), where either both, the syllable distance features and the clause length Syn feature are predictive or neither, while sentence and T-unit length are always predictive. This is also linguistically reasonable, since sentences and T-units can be of considerable length while maintaining small middle fields, while clauses can mainly increase size with a growing middle field. Average clause distance ranks lower than syllable distance between LSB and RSB for grade-wise classification, suggesting that within a clause, the size of the middle field in terms of syllables is more predicting than the clause length in general. This does not hold for school-wise classification, where the ranking of the average syllable distance between LSB and RSB might as well be a symptom of the high ranking for clause length. The same might hold for average syllable distance of the first argument and its main verb, which performs high for grade-wise classification but worse than average clause length, too.

In general, the features from TF have relatively high average merits, which is to be expected given the high ranking of both syllable distance ratios in the overall feature ranking. Unlike with the previous feature rankings, school-wise classification has only slightly higher merits than grade-wise classification for the TF features, while the distance between the merits of average sentence and T-unit length match the previous pattern. Also, in terms of ranks the results are the same for both classification experiments, with the syllable distance between LSB and RSB performing better than the syllable distance between first argument and main verb. Interestingly, the rank for syllable distances on both classification tasks seems to be due to the Publisher B and C corpora, since they rank low on the other two Publisher corpora. A comment on the low ranking for non-subject prefields is still pending. The results suggest that non-subject prefields are not suited to classify either grade-wise or school-wise. It might be the case that adverbials are too common in the left sentence periphery to allow any conclusions based on the position of the subject alone.

4.3.4. Deagentivation Patterns

The feature set internal ranking for deagentivation patterns is shown in Table 4.11. Again, the features are more informative for school-wise than for grade-wise classification. For the latter only the ratio of *man* occurrences is informative. This

Rank	Grade-Wise		School-Wise	
	Average Merit ($\pm\sigma$)	Feature	Average Merit ($\pm\sigma$)	Feature
1.	0.009 (± 0.001)	ratio *man* occurrences	0.029 (± 0.002)	coverage deagentivation patterns
2.	0.000 (± 0.000)	ratio participle modifiers	0.026 (± 0.003)	ratio attributive participles
3.	0.000 (± 0.000)	ratio attributive participles	0.007 (± 0.001)	ratio infinitival constructions
4.	0.000 (± 0.000)	coverage deagentivation patterns	0.006 (± 0.004)	ratio quasi passives
5.	0.000 (± 0.000)	ratio quasi passives	0.005 (± 0.002)	ratio *man* occurrences
6.	0.000 (± 0.000)	ratio passives	0.000 (± 0.000)	ratio passives
7.	0.000 (± 0.000)	ratio halfmodal clusters	0.000 (± 0.000)	ratio participle modifiers
8.	0.000 (± 0.000)	ratio *lassen* occurrences	0.000 (± 0.000)	ratio *lassen* occurrences
9.	0.000 (± 0.000)	ratio infinitival constructions	0.000 (± 0.000)	ratio halfmodal clusters

Table 4.11.: Performance of deagentivation pattern features for grade-wise and school-wise classification in terms of information gain

feature is less informative for school-wise classification, where especially coverage of deagentivation patterns and ratio of attributive participles returns high merits, which are by far the two most predictive features. Distant third, fourth and fifth are the ratios of infinitival constructions, quasi passives and *man* occurrences. Interestingly, participle verb modifiers are not predictive. However, they also occurred only 95 times, while attributive participles occurred 429 times in the Full corpus. The only corpus in which both participle types occur comparably often is in the Publisher D corpus, where 93 attributive and 25 verb modifying participles can be found. For this corpus, neither of the two ranks among the best five features and participle verb modifiers still are assigned the lowest rank. However, this result might not be suitable for generalisations: On the one hand, this corpus only allows for grade-wise classification, but for this task neither participle feature was predictive on the Full corpus. Therefore, results for school-wise classification would be necessary to investigate the differences between the two participle features further. On the other hand, the Publisher D corpus differs strongly in its feature ranking from the other corpora, for example it is the only corpus for which no CompNP feature is among the top five features. Instead, only three features seem to be beneficial at all: ratio of *man* occurrences and coverage of deagentivation patterns as well as of cluster types. Yet, those three features have atypically high average merits each. It remains, therefore, unclear whether participle verb modifiers are less beneficial for classification because of their rareness or because the deagentivational aspect of participles is negligible.

4.3.5. Conditionals

As Table 4.12 shows, the feature set describing conditional sentence is least informative. For school-wise classification, none of the five features has an average

Rank	Grade-Wise		School-Wise	
	Average Merit ($\pm\sigma$)	Feature	Average Merit ($\pm\sigma$)	Feature
1.	0.004 (± 0.002)	ratio v1-dann conditionals	0.000 (± 0.000)	v1-v1 conditionals
2.	0.000 (± 0.000)	ratio wenn-dann conditionals	0.000 (± 0.000)	wenn-dann conditionals
3.	0.000 (± 0.000)	ratio v1-v1 conditionals	0.000 (± 0.000)	coverage conditional types
4.	0.000 (± 0.000)	coverage conditional types	0.000 (± 0.000)	v1-dann conditionals
5.	0.000 (± 0.000)	ratio wenn-v1 conditionals	0.000 (± 0.000)	wenn-v1 conditionals

Table 4.12.: Performance of conditional sentence features for grade-wise and school-wise classification in terms of information gain

merit above 0.0, while for grade-wise classification only the ratio of conditional clauses with a V1 induced protasis and an apposis starting with *dann* is at least marginally informative. This low performance is due to the rareness of all four types of conditional sentences, which occur between six and nine times in the Full corpus. So although the feature set is linguistically well motivated, it cannot be used for either of the given corpora due to its marginality in raw numbers.

4.4. Discussion

The results from section 4.2.2 and section 4.2.3 have shown, that the newly implemented, linguistically motivated features of language complexity improve readability classification. As shown in Table 4.2 and Table 4.3, the new feature sets increase the accuracy of the classifier by Hancke (2013). They perform especially well for school-wise classification, increasing accuracy by up to 1.28%. However, they proof also beneficial for grade-wise classification, increasing accuracy by up to 0.52%. This tendency can be observed throughout all classification experiments and nearly all feature rankings. The new feature sets were not only tested on the classifier from Hancke (ibid.), but also on a version already enhanced by linguistically motivated features implemented by Galasso (2014), to test how much improvement can be reached with more linguistically motivated features. The results show, that the new features still slightly improve the classifier's results. However, the increase never exceeds 1%, which suggests that the classifier reaches the limit of possible improvements by linguistically informed features by employing both, the features presented by Galasso (ibid.) and those presented in this thesis. Interestingly the comparison of the new feature sets from Galasso (ibid.) and this thesis show that the former seems to work better on the smaller Publisher corpora, while the new features in this thesis improve accuracy especially on the Full corpus.

As the experiments on the separate feature sets have shown, CompNP, TF and

Deag are the most beneficial features. Especially postnominal attributes as well as the average syllable distance between LSB and RSB are highly predictive for both classification tasks and return stable confusion matrices even as solitary feature sets. In contrast, the variety of deagentivation patterns and the ratio of attributive participles highly beneficial specifically for school-wise classification. This discrepancy between grade-wise and school-wise classification is highly noticeable, since modifying participles are conceptually associated with phrase complexity, that is CompNP and CompVP, as well as deagentivation patterns. It is therefore unexpected that attributive participles are in fact among the least informative features for grade-wise classification. Similarly unpredicted is the low rank for participle verb modifiers on both classification tasks, which seems to be due to their rareness in the corpus. In general, some patterns that were reasoned to be predictive from a theoretical point of view, proofed irrelevant due to their very low volume, such as participle verb modifiers or future 2. The most striking instance of this issue is the Cond feature set: not for a single feature within this set enough instances were given by the corpus to make any use of them. Overall, the observed differences between the expressiveness of the various phrase modifier types proves the detailed analysis of complex NPs and VPs to be more beneficial than the collection of bare modifiers counts.

Aside from the general evaluation of the newly implemented feature sets, the experiments also investigated differences between the publishers following Vajjala (2015). Their influence on both classification tasks is significant: adding the publisher feature to classification on the Full corpus improves accuracy by up to 5.02%. The repeated experiments on the Publisher corpora allow a even deeper insight in the matter. They illustrate, that often, certain characteristics of the Full corpus are apparently only due to the Publisher corpora A and C, which dominate the Full corpus with their linguistic characteristics due to their considerable size. This can lead to undesirable biases, especially since the Publisher corpora differ strongly in their responsiveness to the different features.[2] For example, neither grade-wise classification on the Publisher C corpus (Table 4.4), nor school-wise classification on the Publisher B corpus (Table 4.5) benefit from more linguistically motivated features. In fact, accuracy drops in both cases for any enhancement with features from either Galasso (2014) or this thesis. Most interestingly, however, is the effect

[2] This aspect is only addressed in extracts within this thesis, since a discussion in greater detail would have been beyond the scope of this work. However, the feature rankings for the Publisher corpora were included for interested readers in Appendix C, Tables C.1 to C.7, pages 71ff.

of the features on the Publisher A corpus, because it is the largest corpus and has consequently the largest effect on the Full corpus. It also shows an interesting pattern: Classification on the Publisher A corpus was improved by enhancement with linguistically motivated features. However, classification accuracy was overall significantly lower than for any other corpus, including the Full corpus. This shows in all experiments as well as in the feature rankings and holds especially true for school-wise classification. Yet, it can also be seen for grade-wise classification. These publisher dependant differences are highly remarkable and should be investigated further in future work.

5. Conclusion

In the course of this thesis overall 46 linguistically motivated features grouped to five feature sets have been implemented with high performance in terms of precision, recall and F-score. The features can be grouped in two categories: those enhancing known features for readability classification with more linguistic information and those implementing recent theoretical findings that have, to my knowledge, not yet been implemented for German readability classification. Deagentivational patterns and non-subject prefields are part of the latter category. Belonging to the former category are complex NPs and VPs, which were modelled in thorough linguistic detail. This has, to my knowledge, not been approached for German readability classification before, either. Inference markers of conditional sentences, too, have been analysed in more linguistic detail including entirely non-mediated conditionals. Finally, linguistically more informed length measures have been approached with Topological Field position based syllable distances.

It was found, that it is in fact beneficial to capture complex NPs and VPs beyond the bare number of modifiers, as their expressiveness is subject to considerable inter-modifier fluctuations. Also, features identified by research on the German register of academic language proved highly useful. In contrast, other promising features, such as the mediation of conditional clauses, were not suited to improve the classification accuracy due to their rareness in the corpus.

Aside from actually improving classification accuracy, the detailed linguistic features also allowed to map considerable differences between publishers, with respect to their production of texts suited for varying grade levels and school types. Therefore, the differences could not only be stated in general based on the strong effect of the publisher feature, but also be analysed in further depth based on the feature set performances and feature rankings on the separate publisher corpora.

Overall, the new features investigated in this thesis lead to interesting insights as well as to slight performance improvements and it seems worthwhile to investigate them further in future work: After the incorporation of complex verbs and deverbal nominalisations as additional features has been approached as a next step, which, unfortunately, could not be completed in time to be included in this thesis, it is planed in the medium term to apply the new features – less tense and passive voice features – to the domain of proficiency assessment, as several of the features can be argued to be beneficial for this domain as well. In this context, not only classification but also clustering based on the new features seems promising for further insights on the linguistic properties of proficiency levels. This thesis, therefore, may be thought of as basis for further work, rather than as a closed case.

Bibliography

Aluísio, Sandra Maria & Gasperin, Caroline. 2010. “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*. Association for Computational Linguistics. Los Angeles, California.

Aluísio, Sandra Maria et al. 2010. “Readability Assessment for Text Simplification”. In: *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics. Los Angeles, California.

Axel, Katrin. 2002. Zur Diachronen Entwicklung der Syntaktischen Integration Linksperipherer Adverbialsätze im Deutschen. Ein Beispiel für syntaktischen Wandel? In: *Beiträge zur Geschichte der deutschen Sprache und Literatur* 124 (1), pp. 1–43.

Ballestracci, Sabrina. 2010. Der Erwerb von Verbzweitsätzen mit Subjekt im Mittelfeld bei italophonen DaF-Studierenden. Erwerbsphasen, Lernschwierigkeiten und didaktische Implikationen. In: *Linguistik online* 41 (1), p. 26.

Behagel, Otto. 1909. Beziehung zwischen Umfang und Reihenfolge von Satzgliedern. In: *Indogermanische Forschungen* 25, pp. 110–142.

Behagel, Otto. 1930. Von deutscher Wortstellung. In: *Zeitschrift für Deutschkunde* 44, pp. 81–89.

Biber, Douglas & Gray, Bethany. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. In: *Journal of English for Academic Purposes* 9, pp. 2–20.

Bohnet, Bernd. 2010. “Top accuracy and fast dependency parsing is not a contradiction”. In: *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics. Beijing, China, pp. 89–97.

Botel, Morton & Granowsky, Alvin. 1972. Syntactic Complexity Formula. In: *A Formula for Measuring Syntactic Complexity: A Directional Effort*, pp. 320–375.

Brants, Thorsten, Skut, Wojciech & Uszkoreit, Hans. 2003. Syntactic annotation of a German newspaper corpus. In: *Treebanks*. Springer, pp. 73–87.

Chall, Jeanne Sternlicht & Dale, Edgar. 1995. *Readability Revisted: The New Dale-Chall Readability Formula*. Brookline Books.

Chang, Chih-Chung & Lin, Chih-Jen. 2013. *A Library for Support Vector Machines*. Tech. rep. National Taiwan University.

Cheung, Jackie Chi Kit & Penn, Gerald. 2009. "Topological Field Parsing of German". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Vol. 1. 1. Association for Computational Linguistics. Suntec, Singapore, pp. 64–72.

Clahsen, Harald, Meisel, Jürgen & Pienemann, Manfred. 1983. *Deutsch als Zweitsprache. Der Spracherwerb ausländischer Arbeiter.* Tübingen, Germany: Narr.

Crossley, S. A. & McNamara, Danielle S. 2011. Understanding expert ratings of essay quality: Coh-metrix analyses of first and second language writing. In: *nternational Journal of Continuing Engineering Education and Life-Long Learning* 21 (2/3), pp. 170–191.

Crossley, Scott A. et al. 2010. Predicting lexical proficiency in language learner texts using computational indices. In: *Language Testing* 20 (10), pp. 1–20.

Dale, Edgar & Chall, Jeanne S. 1948. A Formula for Predicting Readability. In: *Educational research bulletin; organ of the College of Education* 27 (1), pp. 11–28.

D'Alessandro, Donna M., Kingsley, Peggy & Johnson-West, Jill. 2001. The Readability of Pediatric Patient Education Materials on the World Wide Web. In: *Archives of pediatrics and adolescent medicine* 155 (7), pp. 807–812.

Davidson, Donald. 1967. The Logical Form of Action Sentences. In: *The Logic of Decision and Action*. Ed. by N. Resher. University of Pittsburgh Press, pp. 81–95.

Dell'Ortella, Felice, Montemagni, Simonetta & Venturi, Giulia. 2011. "READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification". In: *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics. Edinburgh, Scotland, UK, pp. 73–83.

den Dikken, Marcel & Singhapreecha, Pornsiri. 2004. Complex Noun Phrases and Linkers. In: *Syntax* 7 (1), pp. 1–54.

DuBay, William H. 2004. The Principles of Readability. In: *Online Submission*.

DuBay, William H. 2006. The Classic Readability Studies. In: *Online Submission*.

Duden (Gr). 2009. *Deutsche Grammatik*. Ed. by Ursula Hoberg & Rudolf Hoberg. 4th ed. Vol. 4. Der kleine Duden. Berlin, Germany: Dudenverlag.

Fabricius-Hansen, Cathrine. 2014. *Vorangestellte Attribute und Relativsätze im Deutschen: Wettbewerb und Zusammenspiel*.

Feng, Lijun & Jansche, Martin. 2010. "A Comparison of Features for Automatic Readability Assessment". In: *Coling 2010: Poster Volume*. Beijing, China, pp. 276–284.

François, Thomas & Fairon, Cédrik. 2012. "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*. Association for Computational Linguistics. Jeju Island, Korea, pp. 466–477.

Freda, Margaret Comeford. 2005. The readability of American Academy of Pediatrics patient education brochures. In: *Journal of Pediatric Health Care* 19 (3), pp. 151–156.

Frey, Werner. 2004. The grammar-pragmatics interface and the German prefield. In: *Sprache und Pragmatik* 52.

Frey, Werner. 2005. Zur Syntax der linken Peripherie im Deutschen. In: *Deutsche Syntax: Empirie und Theorie* 46.

Freywald, Ulrike. 2013. Uneingeleiteter V1- und V2-Satz. In: *Satztypen des Deutschen*. Ed. by Jörg Meibauer, Markus Steinbach & Hans Altmann. Berlin, New York: de Gryuter.

Gal, Iddo & Prigat, Ayelet. 2005. Why organizations continue to create patient information leaflets with readability and usability problems: an exploratory study. In: *Health Education Research* 20 (4), pp. 485–493.

Galasso, Sabrina. 2014. *Exploring Textual Cohesion Characteristics for German Readability Classification*. B.A. Thesis.

Gallmann, Peter & Lindauer, Thomas. 1994. Funktionale Kategorien in Nominalphrasen. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur* 116 (1), pp. 1–27.

Graber, Mark A., Roller, Cathy M. & Kaeble, Betsy. 1999. Readability levels of patient education material on the World Wide Web. In: *The Journal of family practice* 48 (1), pp. 58–61.

Graesser, Arthur C. et al. 2004. Coh-Metrix: Analysis of text on cohesion and language. In: *Behaviour Research Methods, Instruments, and Computers* 36 (2), pp. 193–202.

Hancke, Julia. 2013. "Automatic Prediction of CERF Proficiency Levels Based on Linguistic Features of Learner Language". MA thesis. Eberhard Karls Universität Tübingen.

Hancke, Julia, Vajjala, Sowmya & Meurers, Detmar. 2012. "Readability Classification for German using lexical, syntactic and morphological features". In: *Proceedings of COLING*. Mumbai, pp. 1063–1080.

Haufe, Elisabetta Terrasi. 2004. *Der Schulerwerb von Deutsch als Fremdsprache. Eine empirische Untersuchung am Beispiel der italienischsprachigen Schweiz*. Tübingen, Germany: Niemeyer.

Heilman, Michael et al. 2007. "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts". In: *Proceedings of NAACL HLT 2007*. Association for Computational Linguistics. Rochester, NY, pp. 460–467.

Henning, Mathilde & Niemann, Robert. 2013. Unpersönliches Schreiben in der Wissenschaft: Eine Bestandsaufnahme. In: *Informationen Deutsch als Fremdsprache* 4 (439–455).

Hoberg, Ursula. 1981. ie Wortstellung in der geschriebenen deutschen Gegenwartssprache. In: *Heutiges Deutsch. Linguistische Grundlagen. Forschungen des Instituts für deutsche Sprache*. Vol. 10. München, Germany: Max Hueber Verlag.

Höhle, Tilman N. 1986. Der Begriff "Mittelfeld": Anmerkungen über die Theorie der topologischen Felder. In: *Akten des Siebten Internationalen Germanistenkongresses 1985*, pp. 329–340.

Kaiser, Dorothee. 2002. *Wege zum wissenschaftlichen Schreiben: Eine kontrastive Untersuchung zu studentischen Texten aus Venezuela und Deutschland*. Tübingen, Germany: Stauffenburg.

Kincaid, J. P. et al. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Research Branch Report 8-75. Millington, TN: Naval Technical Training Command.

König, Ekkehard & van der Auwera, Johan. 1988. Clause integration in German and Dutch conditionals, concessive conditionals and concessives. In: *Clause Combining in Grammar and Discourse*. Ed. by John Haiman & Sandra A. Thompson. Vol. 18. John Benjamins Publishing, pp. 101–133.

Leacock, Claudia et al. 2014. *Automated Grammatical Error Detection for Language Learners*. Ed. by Graeme Hirst. Morgan & Claypool Publishers.

Levy, Roger & Andrew, Galen. 2006. "Tregex and Tsurgeon: tools for querying and manipulating tree data structures". In: *5th International Conference on Language Resources and Evaluation*.

Lötscher, Andreas. 2006. Linksperiphaere Adverbialsätze in der Geschichte des Deutschen. Pragmatische Aspekte eines grammatischen Wandels. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur* 127 (3), pp. 347–376.

Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. In: *International Journal of Corpus Linguistics* 15 (4), pp. 474–496.

Lu, Xiaofei. 2011. The relationship of lexical richness to the quality of esl learners' oral narratives. In: *The Modern Languages Journal* 96 (2), pp. 190–208.

Lu, Xiaofei & Ai, Haiyang. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. In: *Journal of Second Language Writing* 29, pp. 16–27.

Maienborn, Claudia. 1999. Situationsbezug und die Stadien/Individuen-Distinktion bei Kopula-Prädikativ-Konstruktionen. In: *ZAS Papers in Linguistics* 14, pp. 41–64.

Maienborn, Claudia. 2001. On the Position and Interpretation of Locative Modifiers. In: *Natural Language Semantics* 9 (2), pp. 191–240.

McNamara, Danielle S. et al. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. In: *Cognition and instruction* 14 (1), pp. 1–43.

McNamara, Danielle S. et al. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Camebridge University Press.

Müller, Stefan. 2005. Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. In: *Linguistische Berichte* 203, pp. 297–330.

Müller, Stefan. 2010. German: A Grammatical Sketch. In: *Syntax – Ein internationales Handbuch zeitgenössischer Forschung*. Ed. by Artemis Alexiadou & Tibor Kiss. Berlin, Germany: de Gryuter.

Nietzio, A., Scheer, B. & Bühler, C. 2012. "How Long Is a Short Sentence? – A Linguistic Approach to Definition and Validation of Rules for Easy-to-Read Material". In: *13th International Conference on Computers Helping People with Special Needs*, pp. 369–376.

Payne, John & Berlage, Eva. 2014. Genitive variation: The niche role of the oblique genitive. In: *English Language and Linguistics* 18 (2), pp. 331–360.

Polenz, Peter von. 1981. Über die Jargonisierung von Wissenschaftssprache und wider die Deagentivierung. In: *Wissenschaftliche Beiträge zur Methodologie, the-*

oretischen Fundierung und Deskription. Ed. by Theo Bungarten. München, Germany: Fink, pp. 85–110.

Rafferty, Anna & Manning, Christopher D. 2008. "Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines". In: *ACL Workshop on Parsing German*.

Reis, Marga. 1986. Die Stellung der Verbargumente im Deutschen. Stilblüten zum Grammatik-Pragmatik-Verhältnis. In: *Sprache und Pragmatik. Lunder Symposien 1986*. Ed. by Inger Rosengren. Stockholm, Sweden: Akmqvist and Wiskell International.

Rizzi, Luigi. 2004. On the Form of Chains: Criterial Positions and ECP Effects. In: *Current Studies in Linguistics Series* 42, pp. 97–123.

Roelcke, Thorsten. 2010. *Fachsprachen*. 3rd ed. Berlin, Germany: Schmidt.

Rosenbach, Anette. 2014. English genitive variation – the state of the art. In: *English Language and Linguistics* 18 (2), pp. 215–262.

Schlömer, Anne. 2012. Interkulturelle Aspekte der Wissenschaftskommunikation am Beispiel der Textsorte Wissenschaftlicher Aufsatz. In: *Professional Communication and Translation Studies* 5 (1-2), pp. 48–64.

Schlömer, Anne. 2013. *Erweiterte Nominalgruppen als Merkmal von Wissenschaftssprache. Eine Analyse in Schülertexten und Lehrbüchern*.

Schlotthauer, Susan. 2006. Deutsches Präpositionalattribut und ungarisches Lokalkasus- und Postpositionalattribut. In: *Deutsche Grammatik im europäischen Dialog*, pp. 1–10.

Seeker, Wolfgang & Kuhn, Jonas. 2012. "Making Ellipses Explicit in Dependency Conversion for a German Treebank". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. 3132–3139. Istanbul, Turkey.

Skut, Wojciech et al. 1997. "An annotation scheme for free word order languages". In: *Proceedings of the Fith Conference on Applied Natural Language*. Association for Computational LinguisticsA. Washington, D.C., pp. 88–95.

Telljohann, Heike et al. 2004. "The TüBa-D/Z treebank: Annotating German with a context-free backbone". In: *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 2229–2235.

Thorndike, E. L. 1921. Word Knowledge in the Elementary School. In: *Teachers College Record* 28 (5), pp. 334–370.

Thorndike, E. L. & Lorge, I. 1944. The Teacher's Word Book of 30,000 Words. In: *New York: Teacher's College, Columbia University*.

Todirascu, Amalia et al. 2013. Coherence and cohesion for the assessment of text readability. In: *Natural Language Processing and Cognitive Science* 11, pp. 11–19.

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

Vajjala, Sowmya & Meurers, Detmar. 2012. "On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition". In: *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*. Vol. 7. Association for Computational Linguistics. Montréal, Canada, pp. 163–173.

Vendler, Zeno. 1967. Linguistics in Philosophy. In: Ithaca, New York: Cornell University Press. Chap. 4.

Vogel, M. & Washburne, C. 1928. An Objective Method of Determining Grade Placement of Children's Reading Material. In: *The Elementary School Journal* 28, pp. 373–381.

von der Brück, Tim. 2007. "A Semantically Oriented Readability Checker for German". In: *Proceedings of the 3rd Language and Technology Conference*. Poznań, Poland, pp. 270–274.

von der Brück, Tim. 2008. A Readability Checker with Supervised Learning Using Deep Indicators. In: *Informatica* 32, pp. 429–435.

Witten, Ian, Frank, Eibe & Hall, Mark A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.

Wolfram, Walt. 2006. Variation and language, an overview. In: *Encyclopedia of Languages and Linguistics*, pp. 333–340.

Zeman, Jaromir. 1992. Zur Normalfolge im Mittelfeld. In: *Brünner Beiträge zur Germanistik und Nordistik* 8, pp. 7–15.

Zimmermann, Ilse. 1999. Partizip II - Konstruktionen des Deutschen als Modifikatoren. In: *ZAS Papers in Linguistics* 14, pp. 123–146.

A. Appendix: List of Abbreviations

Glossary

AP adjective phrase. 20, 23

CP complementizer phrase. 30

DP determiner phrase. 19

DWDS Digitales Wörterbuch der deutschen Sprache. 9, 26, 27

LSB left sentence bracket. 26, 30, 31, 45, 50, 51, 53

NLP Natural Language Processing. 15, 23

NP noun phrase. 8, 11, 19, 20, 19, 21, 22, 23, 25, 28, 32, 43, 47, 53, 56

POS Part of Speech. 20, 23, 25, 26, 28, 33

PP prepositional phrase. 20, 21, 25, 28

RSB right sentence bracket. 26, 30, 31, 45, 50, 51, 53

SLA Second Language Acquisition. 10, 11, 12, 14

SMO Sequential Minimal Optimization. 38

SVM Support Vector Machines. 38

VC verb complex. 25, 26, 31

VP verb phrase. 8, 11, 19, 20, 25, 28, 32, 49, 53, 56

WEKA Waikato Environment for Knowledge Analysis. 38

B. Appendix: Feature Performance on Reference Corpus

Counted	Feature set	FP	TP	FN	Precision	Recall	F-Score
# of adjectival verb modifiers	CompVP	0	36	1	1.000	0.973	0.986
# of adverbial verb modifiers	CompVP	0	78	2	1.000	0.975	0.987
# of appositions or parenthesis	CompNP	0	6	21	1.000	0.222	0.363
# of arg1-Vfin-instances	TF	4	49	0	0.925	1.000	0.961
# of attributive participle 1	CompNP	0	4	0	1.000	1.000	1.000
# of attributive participle 2	CompNP	4	21	4	0.840	0.840	0.840
# of clausal noun modifiers	CompNP	0	14	0	1.000	1.000	1.000
# of comparative noun modifiers	CompNP	0	9	2	1.000	0.818	0.900
# of C/FIN-VC instances	TF	0	65	3	1.000	0.956	0.978
# of determiners	CompNP	0	359	1	1.000	0.997	0.998
# of finite verbs	CompVP	0	223	2	1.000	0.991	0.995
# of futur1	CompVP	0	1	0	1.000	1.000	1.000
# of futur2	CompVP	0	0	0	/	/	/
# of half-modal clusters	CompVP	0	0	0	/	/	/
# of identifiable prefields	TF	0	178	0	1.000	1.000	1.000
# of infinitival constructions	Deag	2	39	0	0.951	1.000	0.975
# of inflected main verbs	CompVP	0	113	0	1.000	1.000	1.000
# of *lassen* instances	Deag	0	3	0	1.000	1.000	1.000
# of main verb sub clusters	CompVP	0	12	1	1.000	0.923	0.960
# of *man* instances	Deag	0	2	0	1.000	1.000	1.000
# of modal verb sub clusters	CompVP	1	15	0	0.938	1.000	0.968
# of non canonic prefields	TF	0	24	1	1.000	0.960	0.980
# of NPs	CompNP	4	672	0	0.994	1.000	0.997
# of participle 1 verb modifiers	CompVP	0	3	0	1.000	1.000	1.000
# of participle 2 verb modifiers	CompVP	0	0	0	/	/	/
# of passiv	Deag	0	28	2	1.000	0.933	0.965
# of past perfect	CompVP	0	1	0	1.000	1.000	1.000
# of phi sub clusters	CompVP	0	48	0	1.000	1.000	1.000
# of phrasal postnominal modifiers	CompNP	0	137	1	1.000	0.993	0.996
# of possessive noun modifiers	CompNP	0	61	1	1.000	0.984	0.992
# of prepositional verb modifiers	CompVP	0	242	3	1.000	0.988	0.994
# of prepositions	FunctPP	0	258	0	1.000	1.000	1.000
# of prenominal attributive APs	CompNP	1	201	1	0.995	0.995	0.995
# of quasi passive constructions	Deag	0	4	0	1.000	1.000	1.000
# of simple past	CompVP	1	17	0	0.944	1.000	0.971
# of simple perfect	CompVP	0	14	0	1.000	1.000	1.000
# of simple present	CompVP	0	176	1	1.000	0.994	0.997
# of subjects	TF	0	200	0	1.000	1.000	1.000
# of syllables between arg1 and Vfin	TF	4	477	9	0.992	0.981	0.986
# of syllables between C/FIN and VC	TF	0	604	10	1.000	0.984	0.992
# of v1-*dann* conditionals	Cond	0	0	0	/	/	/
# of v1-v1 conditionals	Cond	0	1	1	1.000	0.500	0.667
# of V2 clusters	CompVP	0	15	2	1.000	0.882	0.937
# of V3 clusters	CompVP	0	2	0	1.000	1.000	1.000
# of V4 clusters	CompVP	0	0	0	/	/	/
# of V5 clusters	CompVP	0	0	0	/	/	/
# of V6 or higher clusters	CompVP	0	0	0	/	/	/
# of verb particles	CompVP	0	19	6	1.000	0.76	0.864
# of VPs	CompVP	0	291	2	1.000	0.993	0.996
# of *wenn-dann* conditionals	Cond	0	0	0	/	/	/
# of *wenn*-V1 conditionals	Cond	1	5	0	0.833	1.000	0.909
Counted	Feature set	FP	TP	FN	Precision	Recall	F-Score

Table B.1.: Performance of all countings on Reference corpus including false positives (FP), true positives (TP) and false negatives (FN).

C. Appendix: Feature Ranking on Publisher Corpora

Rank	Feature Set	Average Merit (±σ)	Feature
–	Syn	0.023 (± 0.002)	average sentence length
1.	Comp NP	0.022 (± 0.001)	ratio postnominal modifier
–	Syn	0.018 (± 0.009)	average T-unit length
2.	Comp VP	0.000 (± 0.000)	ratio periphrastic tense
3.	Comp VP	0.000 (± 0.000)	ratio present perfect
4.	Comp VP	0.000 (± 0.000)	ratio future 1
5.	Comp VP	0.000 (± 0.000)	coverage modifier types

Table C.1.: Five best features for grade-wise classification from all feature sets in terms of information gain for publisher A.

Rank	Feature Set	Average Merit (±σ)	Feature
1.	Comp NP	0.017 (± 0.002)	ratio postnominal modifiers
2.	Deag	0.011 (± 0.002)	coverage deagentivation patterns
3.	Comp VP	0.000 (± 0.000)	ratio present perfect
4.	Comp VP	0.000 (± 0.000)	ratio future 1
5.	Comp VP	0.000 (± 0.000)	ratio periphrastic tenses

Table C.2.: Five best features for school-wise classification from all feature sets in terms of information gain for publisher A.

Rank	Feature Set	Average Merit (±σ)	Feature
–	Syn	0.056 (± 0.006)	average sentence length
–	Syn	0.054 (± 0.005)	average T-unit length
–	Syn	0.045 (± 0.004)	average clause length
1.	VC	0.042 (± 0.004)	average syllable distance LSB/RSB
2.	Deag	0.041 (± 0.004)	ratio *man* occurrences
3.	Comp NP	0.043 (± 0.006)	ratio postnominal modifiers
4.	CompVP	0.000 (± 0.000)	ratio periphrastic tense
5.	CompVP	0.000 (± 0.000)	ratio future 1

Table C.3.: Five best features for grade-wise classification from all feature sets in terms of information gain for publisher B.

Rank	Feature Set	Average Merit (±σ)	Feature
–	Syn	0.241 (± 0.009)	average sentence length
–	Syn	0.135 (± 0.016)	average T-unit length
1.	VC	0.023 (± 0.002)	average syllable distance LSB/RSB
2.	Comp NP	0.020 (± 0.001)	ratio postnominal modifiers
3.	Comp NP	0.019 (± 0.001)	ratio possessive modifiers
4.	TF	0.017 (± 0.002)	average syllable distance arg1 to main verb
5.	Comp NP	0.011 (± 0.001)	ratio prenominal modifiers

Table C.4.: Five best features for school-wise classification from all feature sets in terms of information gain for publisher B.

Rank	Feature Set	Average Merit (±σ)	Feature
–	Syn	0.065 (± 0.008)	average clause length
–	Syn	0.059 (± 0.009)	average sentence length
1.	TF	0.056 (± 0.010)	average syllable distance LSB/RSB
2.	Comp NP	0.059 (± 0.012)	ratio prenominal modifiers
–	Syn	0.055 (± 0.004)	average T-unit length
3.	Comp NP	0.049 (± 0.011)	ratio attributive participles
4.	Comp NP	0.045 (± 0.004)	ratio postnominal modifiers
5.	Deag	0.041 (± 0.010)	ratio *man* occurrences

Table C.5.: Five best features for grade-wise classification from all feature sets in terms of information gain for publisher C.

Rank	Feature Set	Average Merit (±σ)	Feature
–	Syn	0.131 (± 0.007)	average sentence length
–	Syn	0.084 (± 0.007)	average T-unit length
–	Syn	0.082 (± 0.006)	average clause length
1.	Comp NP	0.077 (± 0.011)	ratio prenominal modifiers
2.	Comp NP	0.036 (± 0.004)	ratio determiners
3.	TF	0.034 (± 0.005)	average syllable distance LSB/RSB
4.	TF	0.029 (± 0.003)	average syllable distance arg1 to main verb
5.	Deag	0.029 (± 0.004)	coverage deagentivation patterns

Table C.6.: Five best features for school-wise classification from all feature sets in terms of information gain for publisher C.

Rank	Feature Set	Average Merit ($\pm\sigma$)	Feature
1.	DEAG	0.097 (± 0.023)	ratio # *man*
2.	DEAG	0.054 (± 0.005)	coverage deagentivation. patterns
3.	COMP VP	0.044 (± 0.007)	coverage cluster types
4.	COMP VP	0.000 (± 0.000)	ratio periphrastic. tense
5.	COMP VP	0.000 (± 0.000)	coverage modifier types

Table C.7.: Five best features for grade-wise classification from all feature sets in terms of information gain for Publisher D.