Readability Classification for German using lexical, syntactic, and morphological features

Julia Hancke *Sowmya Vajjala* *Detmar Meurers*
Seminar für Sprachwissenschaft, Universität Tübingen
{jhancke,sowmya,dm}@sfs.uni-tuebingen.de

Abstract

We investigate the problem of reading level assessment for German texts on a newly compiled corpus of freely available *easy* and *difficult* articles, targeted at adult and child readers respectively. We adapt a wide range of syntactic, lexical and language model features from previous research on English and combined them with new features that make use of the rich morphology of German. We show that readability classification for German based on these features is highly successful, reaching 89.7% accuracy, with the new morphological features making an important contribution.

Title and Abstract in German

Lesbarkeitsklassifizierung für das Deutsche mit lexikalischen, syntaktischen und morphologischen Merkmalen

Wir untersuchen das Problem der Lesbarkeitsklassifizierung für deutsche Texte anhand eines neuen Korpus frei zugänglicher Artikel, die einerseits Erwachsene und andererseits Kinder als Zielgruppe haben. Wir adaptieren eine Vielzahl syntaktischer, lexikalischer und *language model* Merkmale aus der englischen Lesbarkeitsforschung und kombinierten sie mit neuen Merkmalen, die sich die ausgeprägte Morphologie des Deutschen zu Nutze machen. Wir zeigen, dass diese Merkmale sehr erfolgreich dazu eingesetzt werden können, deutsche Texte nach ihrer Lesbarkeit zu klassifizieren. In unseren Experimenten erreicht die Klassifikation eine Genauigkeit von 89,7%, wozu die neuen morphologischen Merkmale einen wichtigen Beitrag leisten.

Keywords: Readability, Complexity, Simplification, Second Language Acquisition, Proficiency.

Keywords in German: Lesbarkeit, Komplexität, Textvereinfachung, Zweitspracherwerb, Sprachniveau.

1 Introduction

The last decade has seen an increasing interest in building automatic readability assessment systems. Such systems can help users in finding texts that they can understand, for example by identifying appropriate texts from the huge number of documents available on the web (Bennöhr, 2005; Miltsakaki, 2009; Ott and Meurers, 2010; Collins-Thompson, 2011). This is particularly relevant for first or second language learners as well as for people with intellectual disabilities. Readability classification systems can also be used as a starting point in identifying targets for text simplification, with the goal of ensuring that information can be accessed and understood by a broader audience. The need for such applications is likely to increase given the increasing relevance of information from the web for everyday life.

While early research on readability assessment derived readability formulae from superficial language properties, most current research takes advantage of natural language processing tools to analyze the texts. The resulting features are combined for classification using machine learning. While syntactic, lexical, language model and discourse features have been examined extensively in readability classification, the influence of morphological indicators has received only little attention. This may also be due to the fact that most of the readability research has focused on English. However, there is some recent interest in automatic readability assessment for other languages such as French, German, Italian and Portuguese.

In this paper, we present and evaluate a readability classification approach for German, as a first step towards our overall goal of identifying targets for text simplification. Since there are no existing German corpora that met our needs, we created a two-level readability corpus collected from publicly accessible websites. On this basis, we explore a wide range of syntactic, lexical and language model features derived from previous research on English. In addition, as German has a rich morphology, morphological features may also provide valuable information on the reading level. Hence, we devised a new set of features based on the inflectional and derivational morphology of German. We then conducted classification experiments to examine how well the different feature groups work as indicators of the reading level. We also compared the performance of these feature groups in isolation as well as in combination with each other. To summarize, in this paper, we show that the features used in English readability research can be successfully applied to German, and that the addition of German language-specific morphological features will improve the classification accuracy.

The paper is organized as follows: Section 2 summarizes related work in the field of readability assessment. Section 3 provides information about the origin and nature of our dataset. Section 4 introduces our approach to readability classification including the features we used. Section 5 describes our experimental setup and the results. We conclude this paper with a discussion of the results and directions of future work.

2 Related Work

Research on English readability assessment has a long history spanning several decades (DuBay, 2006). Many traditional readability formulae, such as the Flesch-Kincaid score (Kincaid et al., 1975), relied on easy to calculate approximations of syntactic or lexical complexity, such as number of characters per word or average sentence length. Other early approaches, like the Dale-Chall formula (Chall and Dale, 1995), approximated semantic complexity by using word frequency lists. More recent approaches benefit from advances in natural language processing and machine learning. Si and Callan (2001) and Collins-Thompson and Callan (2004) used a unigram model for readability classification. Heilman et al. (2007, 2008) combined unigram models with grammatical features and trained machine learning models for readability assessment. Their aim in the context of the *REAP* project (`http://reap.cs.cmu.edu`) was to retrieve reading material for language learners.

Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009) combined traditional features with syntactic parse tree features and n-gram language models. Their work is based on the *Weekly Reader*, an educational newspaper consisting of articles at four reading levels. Feng (2010) also used the *Weekly Reader* dataset to build classification models with several discourse features alongside parse features, language model features, and traditional features. The discourse features are motivated by the cognitive processes involved in text understanding and underline their focus on finding appropriate texts for people with mild intellectual disabilities. Crossley and McNamara (2011) and the overall *CohMetrix* project (`http://cohmetrix.memphis.edu`) also explored a wide range of cognitively grounded features related to text cohesion and coherence. Vajjala and Meurers (2012) showed that reading level assessment can benefit from Second Language Acquisition (SLA) research. They enriched the lexical and syntactic features from previous approaches by using features derived from measures of the language proficiency of learners.

Readability research on English has ignored morphological features to a large extent. However, with recent interest in readability assessment for languages other than English, the use of features which are relevant for other languages is gaining some prominence. Research on Italian and French readability is taking advantage of the rich verbal morphology of these languages. Dell'Orletta et al. (2011) worked with a corpus of Italian newspaper text at two different reading levels. They used a mixture of traditional, morpho-syntactic, lexical and syntactic features for building a two class readability model for Italian. Among others, their feature set included verbal mood based features, which relied on the rich verbal morphology of Italian. François and Fairon (2012) built their French readability classification model using a text book corpus designed for adult learners of French. They also considered verb tense and mood based text difficulty features along with several other features. Readability assessment was also studied for Portuguese using various lexical, syntactic, discourse and language modeling features derived from English research (dos Santos Marujo, 2009; Aluisio et al., 2010). Lau (2006) utilized the nature of the Chinese script to form several sub-character and character level features in addition to the common word and sentence level features for Chinese readability classification.

The only published research on German readability assessment that we know of is the *DeLite* readability checker (Vor der Brück and Hartrumpf, 2007; Vor der Brück et al., 2008a,b). *DeLite* was built using a human annotated corpus of 500 texts from the municipal domain, such as city ordinances. It was classified into ten levels of difficulty. The corpus includes mostly legal texts and is generally at a higher level of reading difficulty compared to ordinary texts. *DeLite* makes use of a comprehensive set of features that aim to capture readability at lexical, syntactic, semantic and discourse levels. They also considered some morphological indicators, such as the number of nouns that are derived from adjectives or verbs, the complexity of compounds, and the number of acronyms.

Due to the lack of multi-level graded corpora for languages other than English, researchers often built readability models from freely available collections of two or three classes collected from the web. Dell'Orletta et al. (2011), Aluisio et al. (2010), and Klerke and Søgaard (2012) report on creating and experimenting with such corpora in Italian, Portuguese and Danish respectively. Napoles and Dredze (2010) performed a similar experiment for English with a corpus built from Wikipedia and Simple Wikipedia. Incidentally, all of these groups worked on readability assessment in the context of text simplification.

3 Data: The *GEO-GEOlino* Corpus

Research on the readability classification of English texts has often used the *Weekly Reader* as a gold standard. An established resource that could be compared to this dataset does not exist for German. The only existing German readability corpus is the one collected for the readability checker *DeLite*, but as discussed in the previous section, it is a domain specific collection, consisting mostly of legal texts from the municipal domain. Hence, we created our own two class corpus (*easy* vs. *difficult*) with articles collected from the web, based on the assumption that texts written for children are easier to read than those for adults.

We crawled articles from the websites of two related German magazines, *GEO* (`http://www.geo.de`) and *GEOlino* (`http://www.geolino.de`), published by Gruner & Jahr. *GEO* is a monthly magazine containing articles in the domains of nature, culture and science. *GEOlino* is a magazine on similar topics by the same publisher, but it is targeted at children from age 8–14. *GEOlino* it is not a simplified version of *GEO*; the content is specifically created for child readers.

Table 1 shows the distribution of the articles we crawled across all topics. The overall corpus we collected consists of 4603 articles from both websites.[1]

	GEOlino			**GEO**		
Topics:	**Num. tokens**	**Num. sentences**	**Num. articles**	**Num. tokens**	**Num. sentences**	**Num. articles**
Nature	189 004	13 976	321	877 920	54 300	1 459
Human	412 769	33 497	901	443 221	29 482	662
Technology	57 819	4 356	83	204 891	12 674	392
Culture	–	–	–	442 888	30 748	463
Creative	169 800	12 354	322	–	–	–
Total	829 392	65 183	1 627	1 968 920	127 204	2 976

Table 1: Composition of the *GEO-GEOlino* corpus

For the experiments discussed in this paper, we randomly selected an equal number of documents from each of the topics that existed in both *GEO* and *GEOlino*: 321 articles from Nature, 662 from Human and 83 from Technology. We cleaned the data obtained from the web by removing all html markup, meta data, and non-text content. We also eliminated duplicate articles. We then tagged the corpus using a Java interface[2] to the RFTagger (Schmid and Laws, 2008), a statistical tagger that provides a fine grained morphological analysis. The tagged articles, mapped to the Stuttgart-Tübingen Tagset (STTS) for German, were then parsed with the Stanford Parser for German (Rafferty and Manning, 2008), which comes with a German model trained on NEGRA.[3]

4 Features

We modeled readability using five groups of features: features from traditional readability formulas, lexical features, syntactic features, language model features, and morphological features. For the first three groups, we essentially adapted the English features described by

[1] Contact us by email if you are interested in using this corpus for non-commercial research purposes.

[2] `http://www.sfs.uni-tuebingen.de/~nott/rftj-public/`

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

Vajjala and Meurers (2012) to German. Given the effectiveness of language models in previous work on English (e.g., Petersen and Ostendorf, 2009; Feng, 2010), we included language model features as our fourth group. In the fifth group, we explored new features that encode aspects of the inflectional and derivational morphology and compounding of German.

4.1 Traditional Features (TRAD)

The traditional features group we used includes the average sentence length in words, the average number of characters, and syllables per word. These properties have often been used in traditional readability formulae (e.g., Kincaid et al., 1975). Although they do not analyze readability at a deeper level, they have been popular in the English readability literature for a long time. They also constitute a useful baseline against which to interpret the effect of broader and deeper analysis in current readability classification.

4.2 Lexical Features (LEX)

Vajjala and Meurers (2012) employed the lexical richness measures that were originally developed for judging the proficiency of second language learners (Lu, 2011) for English readability classification. We adapted these features for German and added some further features we considered worth exploring. We added the noun token ratio, the verb token ratio and the verb-noun token ratio as additional lexical features, with the hypothesis that easy documents will include fewer nominalizations compared to difficult documents. Additionally, we added the ratio of *sein* to verbs and *haben* to verbs. The list of all implemented lexical features is shown in Table 2. As the class of lexical words (Lex) mentioned in several of the formulas, we used nouns, adjectives, adverbs, and full verbs – i.e., in terms of the STTS tagset: AD.*, N.*, VV.*.

Lexical Richness Features from SLA	**Other Lexical Features**
Type-Token Ratio	*sein* to Verb Token Ratio
Root Type-Token Ratio	*haben* to Verb Token Ratio
Corrected Type-Token Ratio	Avg. Num. Characters per word
Bilogarithmic Type-Token Ratio	Avg. Num. Syllables per word
Uber Index = $\log(Typ^2)/\log(Tok/Typ)$	Verb Token Ratio = Tok_{Verb}/Tok
Measure of Textual Lexical Diversity (McCarthy, 2005)	Noun Token Ratio
Lexical Density = Tok_{Lex}/Tok	Verb-Noun Token Ratio
Lexical Word Variation = Typ_{Lex}/Tok_{Lex}	
Noun Variation = Typ_{Noun}/Tok_{Lex}	
Adjective Variation, Adverb Variation	
Modifier Variation = $(Typ_{Adj} + Typ_{Adv})/Tok_{Lex}$	
Verb Variation 1 = Typ_{Verb}/Tok_{Verb}	
Verb Variation 2 = Typ_{Verb}/Tok_{Lex}	
Squared Verb Variation 1 = Typ^2_{Verb}/Tok_{Verb}	
Corrected Verb Variation 1 = $Typ_{Verb}/\sqrt{2Tok_{Verb}}$	

Table 2: The lexical features we implemented

4.3 Syntactic Features (SYN)

For the syntactic features, we adapted a range of parse tree based features from English readability assessment research to German. The features include the average parse tree height and the average length and number of NPs, VPs and PPs per sentence (Petersen and Ostendorf, 2009; Feng, 2010). Following Vajjala and Meurers (2012), we also adapted proficiency measures from SLA (Lu, 2010), including various ratios that try to capture level of embedding and coordination, length of production unit, and relationships between specific structures. Table 3 lists all the syntactic features we implemented.

Syntactic Features from SLA	**Other Syntactic Features**
Avg. length of a clause	Num. NPs per sentence
Avg. length of a sentence	Num. VPs per sentence
Avg. length of a T-unit	Num. PPs per sentence
Num. of Clauses per Sentence	Num. VZs per sentence
Num. of T-Units per sentence	Avg. length of a NP
Num. of Clauses per T-unit	Avg. length of a VP
Num. of Complex-T-Units per T-unit	Avg. length of a PP
Dependent Clause to Clause Ratio	Num. Dependent Clauses per sentence
Dependent Clause to T-unit Ratio	Num. Complex-T units per sentence
Co-ordinate Phrases per Clause	Co-ordinate Phrases per sentence
Co-ordinate Phrases per T-unit	Avg. parse tree height
Complex Nominals per Clause	
Complex Nominals per T-unit	
Verb phrases per T-unit	

Table 3: The syntactic features we implemented

Three basic production units, *sentences*, *clauses*, and *T-Units* are relevant for computing these measures. We adapted their definitions from those used in the SLA literature (Lu, 2010). A *sentence* is a group of words delimited with a punctuation mark (period, question mark, exclamation mark, or quotation mark). We implemented sentence splitting using the Stanford Document Processor with the default tokenizer factory.

Clauses for English are characterized as structures that contain a subject and a finite verb (Hunt, 1965). Different from English, German allows subjectless sentences, so we consider all maximal projections headed by a finite verb, as well as elliptical constructions where the finite verb is omitted. One word exclamations such as *Stop!* are excluded. In the parsed data, this notion of a *clause* corresponds to the category S of the NEGRA annotation scheme (Skut et al., 1997).

T-Units are defined as "one main clause plus any subordinate clause or non clausal structure that is attached to or embedded in it" (Hunt 1970, p. 4; cf. Lu, 2010). Only independent clauses (including their dependents) count as a T-unit. Consider, for example, the sentence in (1).

(1) Tom fragt Maria, ob sie die Wahrheit sagt, aber sie antwortet nicht.
Tom asks Maria whether she the truth says but she answers not

'Tom asks Maria, whether she says the truth, but she does not answer.'

Figure 1 shows the parse tree of this example sentence. It includes three clauses (shown by the category S) but only two T-Units (indicated by rectangles). One of the clauses is a dependent clause (underlined) and the T-Unit containing it is therefore a complex T-Unit.

```
(ROOT
  (CS
   (S
     (NP-SB (NE Tom))
     (VVFIN fragt)
     (NP-OA (NE Maria) ($, ,)
       (S (KOUS ob)
         (NP-SB (PPER sie))
         (NP-OA (ART die) (NN Wahrheit))
         (VVFIN sagt))))
   ($, ,) (KON aber)
   (S
     (NP-SB (PPER sie))
     (VVFIN antwortet) (PTKNEG nicht))
   ($. .)))
```

Figure 1: Parse tree of example sentence (1)

At the phrasal level, *coordinated phrases* are defined as coordinated adjective, adverb, noun, and verb phrases. *Verb phrases* include non-finite as well as finite verb phrases. Finally, *complex nominals* are nouns with an adjective, possessive, possessive, prepositional phrase, relative clause, participle, or appositive; nominal clauses, gerunds and infinitives in subject position are also included.

In addition to adapting the patterns and categories for German, we also added VZs ('zu'-marked infinitives) per phrase as a feature. In the so-called incoherent constructions in German (Bech, 1955; Meurers, 2000), the verb always appears in the 'zu'-infinitival form and the phrase it heads is similar in function to a subordinate clause. This naturally is only a coarse approximation of clause-like infinitival constructions given that many verbs selecting 'zu'-infinitivals can alternatively appear in non-clausal, coherent constructions.

On the practical side, TregEx (Levy and Andrew, 2006) was used to extract the patterns that identify the relevant syntactic structures. Lu (2010) created such patterns for English. We modified those patterns to suit the German parse tree structures and categories. As an example, (2a) shows the TregEx pattern for coordinated phrases, for which (2b) is an example structure from our corpus (translating to *parents and children*). (3a) shows the regular expression for complex nominals and (3b) a corpus instance matching that pattern (*a spoiled child*).

(2) a. CAC|CAVP|CNP|CVP
    b. (CNP (NN Eltern) (KON und) (NN Kindern))

(3) a. @NP|NN|NE < S | < @AP | < @PP | < @PP | < ADJA|ORD
    b. (NP (ART ein) (ADJA verwöhntes) (NN Kind))

4.4 Language Model Features

Following Petersen and Ostendorf (2009), we chose a separate data set for training the language models. For building the *easy* language model, we collected 2000 articles from *News4Kids* (`http://www.news4kids.de`), a German website which adapts news found on the web for children. The *difficult* language model was created using 2000 articles from the website of the

German news channel *NTV* (`http://www.n-tv.de`). For these web texts, we followed the same preprocessing routine as explained in Section 3.

Schwarm and Ostendorf (2005) and Feng (2010) suggested that mixed models combining words and parts of speech are more effective for readability assessment than simple word based models. But Petersen and Ostendorf (2009) reached the opposite conclusion. We experimented with both types of models. For preparing the mixed models, we followed the procedure suggested in Petersen and Ostendorf (2009). We trained a bag-of-words classifier with our language modeling dataset (*News4Kids* and *NTV*) and employed Information Gain (Yang and Pedersen, 1997) for feature selection. All words below an empirically determined threshold were replaced by their part of speech.

We used the SRI Language Modeling toolkit (Stolcke, 2002) for training unigram, bigram and trigram models on the *words-only* and *mixed word/part of speech* corpora for each of our two reading levels. This resulted in twelve language models. For all models, we selected Kneyser-Ney smoothing (Chen and Goodman, 1999) as smoothing technique. The perplexity scores from all twelve language models were used as features for building the classifier. These features are listed in Table 4.

Level of Difficulty	**Word Based Model**	**Mixed Model**
Easy	Unigram perplexity	Unigram perplexity
	Bigram perplexity	Bigram perplexity
	Trigram perplexity	Trigram perplexity
Difficult	Unigram perplexity	Unigram perplexity
	Bigram perplexity	Bigram perplexity
	Trigram perplexity	Trigram perplexity

Table 4: The twelve perplexity scores used as language model features

4.5 Morphological Features (Morph)

German has a rich inflectional and derivational morphology. Although not to the same extent as Romance languages, German uses inflectional morphemes to convey a range of grammatical meanings. For example, person and number of a verb are indicated by inflectional endings (e.g., *ich kaufe* [*I buy*], *du kaufst* [*you buy*]). On the nominal side, German has four cases and nouns fall into several different declension paradigms. Case information is sometimes carried by the articles and sometimes by the noun.

German derivational morphology uses various prefixes and suffixes. Nominalizations with a suffix (*regieren* [*govern*] – *Regierung* [*government*]) or without (*laufen* [*to run*] – *der Lauf* [*the run*]) are very common. Compounding is another productive morphological process in German.

Since these three morphological processes may reflect distinctions of relevance for identifying the reading level of a text, we included them in our experiments.

4.5.1 Inflectional Morphology of the Verb (InflV) and the Noun (InflN)

We wanted to investigate if the verbal tense and mood encodes relevant information about a text's difficulty. Hence, we examined a broad range of features related to the inflectional properties of the verb including person, tense, mood and type of verb (finite, non-finite, auxiliary). Additionally, we included the case properties of nouns as features. The case of a nominal argument, for example, reflects the nature and complexity of the subcategorization frame of the head that selects it. The list of all our features related to inflectional morphology is shown in Table 5.

Verb (INFLV)	**Noun** (INFLN)
Num. infinitive Vs / Num. Vs	Num. accusative Ns / Num. Ns
Num. participle Vs / Num. Vs	Num. dative. Ns / Num. Ns
Num. imperative Vs / Num. Vs	Num. genitive Ns / Num. Ns
Num. present tense Vs / Num finite Vs	Num. nominative Ns / Num. Ns
Num. past tense Vs / Num. finite Vs	
Num. 1st person Vs / Num. finite Vs	
Num. 2nd person Vs / Num. finite Vs	
Num. 3rd person Vs / Num. finite Vs	
Num. subjunctive Vs / Num. finite Vs	
Num. finite Vs / Num. Vs	
Num. modal Vs / Num. Vs	
Num. auxiliary Vs / Num. Vs	
Num. Vs / Num. S	

Table 5: The features based on inflectional morphology

4.5.2 Derivational Morphology of the Noun (DERIV)

Two features based on derivational morphology were previously used for German. Vor der Brück et al. (2008b) measured the number of words that are derived from a verb or an adjective. Derived nouns are linguistically more complex than simple nouns. Additionally, derivational suffixes often differ for words which are native German words and lexical items that have come into German from other languages, e.g., *linguist* can be translated into German as either *Lingu**ist*** or as *Sprachwissenschaft**ler***.

We included relatively fine grained derivational properties by taking into account the distribution of a number of individual suffixes. We manually compiled a set of suffixes from a comprehensive overview of native and foreign suffixes (Fleischer and Barz, 1995). For each suffix, we listed all of the different gender and number forms. The counts are accumulated per type of suffix. For example, if there are two instances of *-ismus* and one instance of *-ismen* (the plural form), then a sum of three instances is recorded for *-ismus*. To avoid counting instances of homomorphic nouns as suffixes (e.g., *Ei* [*egg*] vs. suffix *-ei*), we only considered polysyllabic words for the suffix analysis. A list of all the suffixes we used is shown in Table 6.

All occurrences of each suffix (and its forms) were counted per document. Three different ratios can then be generated for each suffix. Let S be the count of a given suffix from Table 6 for a given document, and let T, N and DN be the number of tokens, nouns, and derived nouns in that document. We defined derived noun as a noun which ends in one of the suffixes listed in Table 6. The suffix token ratio (STR), suffix noun ratio (SNR) and suffix derived noun ratio (SDNR) can then be calculated as follows:

- **Suffix Token Ratio (STR)** $= \frac{1}{T}S$
- **Suffix Noun Ratio (SNR)** $= \frac{1}{N}S$
- **Suffix Derived Noun Ratio (SDNR)** $= \frac{1}{DN}S$

suffix	further suffix forms
ant	anten, antin, antinnen
arium	arien
ast	asten, astin, astinnen
at	ate
ator	atoren, atorin, atorinnen
atur	aturen
ei	eien
er	erin, erinnen
ent	ents
enz	enzen
eur	eure, eurin, eurinnen
heit	heiten
ist	isten, istin, istinnen
ion	ionen
ismus	ismen
ität	itäten
keit	keiten
ling	lingen
nis	nisse
schaft	schaften
tum	tümer
ung	ungen
ur	
werk	werke
wesen	

Table 6: List of German derivational suffixes used

For example, consider the sentence in (4) and assume that this is a one sentence document.

(4) Die Regier**ung** muss die Nutz**ung** der landwirtschaftlichen Flächen mit großer
the government has to the use of agricultural areas with great
Präzis**ion** planen.
precision plan

'The government has to plan the use of the agricultural areas with great precision.'

In this example, there are two instances of the suffix *-ung* and one instance of the suffix *-ion* (shown in bold). *Regierung* and *Nutzung* are both nominalizations of native words with native origin, and *Präzision* is of foreign origin. This one sentence document includes twelve tokens, four nouns and three derived nouns, so one obtains STR(ung) = 2/12, SNR(ung) = 2/4 and SDNR(ung) = 2/3. For the suffix *-ion* we get 1/12, 1/4 and 1/3, respectively. The values for all other suffixes are zero. In addition to the specific suffix ratios, we also computed the overall ratio of derived nouns to nouns.

4.5.3 Nominal Compounds (COMP)

Compounds are frequent in German and compounding is a productive mechanism of word formation in German Fleischer and Barz (1995, p. 85). Vor der Brück et al. (2008b) included two compounding features into the classifier for the readability checker *DeLite*: the number of words that form a compound and the number of semantic concepts that are incorporated into a compound. In our approach, we considered the ratio of compound nouns to all nouns and the average number of words in a compound as our compounding based features. For identifying compounds and splitting them into their parts we used JWordSplitter 3.4[4].

5 Experiments and Results

We performed our classification experiments using the *Geo-GeoLino* dataset described in Section 3. We employed the *WEKA* (Hall et al., 2009) implementation of the Sequential Minimal

[4] http://www.danielnaber.de/jwordsplitter

Optimization (SMO) algorithm to create our classification models. For all our classification experiments, we report the overall accuracy after 10-fold cross validation.

We first report the results with language modeling (LM) and the morphological feature groups separately. We then report the performance of each of the other feature groups introduced in Section 4 separately, before turning to combinations.

5.1 Language Modeling

In previous research on the use of language model features for readability assessment, it remained unclear if word based models or mixed word/part of speech models are more effective. While Schwarm and Ostendorf (2005) reported that the mixed models performed better, this was not confirmed by Petersen and Ostendorf (2009). Schwarm and Ostendorf (2005) trained the language models on the same dataset they used for training their classifier, whereas Petersen and Ostendorf (2009) used different datasets for training the language models and the classifier. Feng (2010), who also trained the language models on the same data as the readability classifier, reports that the mixed models outperformed the purely word based models.

As we discussed in section 4.4, we followed Petersen and Ostendorf (2009) in training the language model on a separate data set, to test whether the information picked up about easy vs. difficult texts generalizes across corpora. We trained one classifier on the purely word based models and another one on the mixed model. We combined all perplexity scores from both word based and mixed models. In our experiments, the mixed model (71%) performed only slightly better than the word based model (70.8%). However, an experiment using all twelve perplexity scores as features showed that a combination of mixed and word based models improved classification accuracy (77.6%). So we kept all twelve scores in the feature group for further experiments.

5.2 Morphological Features

We investigated the morphological features in detail given that most of them had not been employed before. In the case of derivational morphology, as described in Section 4.5.2, we calculated three ratios: suffix token ratio, suffix noun ratio and suffix derived noun ratio for each of the 25 derivational suffixes and also calculated the overall ratio of derived nouns to nouns, leading to a total of 76 features. We first performed classification experiments with all three suffix ratio feature subsets separately. The suffix token ratio features performed best (76.6%), followed by suffix noun ratio features (74.4%) and suffix derived noun ratio features (74.0%). However, a combination of all the derivational suffix based features achieved a higher accuracy (78.5%) than the individual subsets. The verbal inflection (InflV) and nominal inflection (InflN) features (Table 5) alone achieved accuracies of 74.3% and 67.2%, respectively. Combining all the inflectional features resulted in a classifier that performed much better, at 79.0%.

Table 7 shows the results for each of our morphological feature groups. The derivational features (Deriv) achieved the highest accuracy followed by the features that use verbal inflection (InflV) and nominal inflection (InflN). The nominal compound features (comp) were the least effective predictors among the morphological groups.

The combined feature set (MORPH) consisting of all four morphological subsets performed with an accuracy 85.4%, which is better than any of the individual subsets. Removing the compounding features from MORPH had no impact on the accuracy.

Feature set	Num. Features	Accuracy
DERIV	76	78.5%
INFLN	4	67.2%
INFLV	13	74.3%
INFLN & INFLV	17	79.0%
COMP	2	56.96%
MORPH	95	85.4%

Table 7: Results for the different types of morphological features

5.3 Most Predictive Features

Apart from the various feature subsets, we also determined the most predictive features using Information Gain. The ten most predictive features (TOP 10) are shown in Table 8. Training a classifier with the TOP10 features resulted in an accuracy of 84.3%.

Feature	**Group**
Avg. Word Length	Lex/Trad
Num. 2nd person Vs / Num. finite Vs	Morph
Num. Syllables Per Word	Lex/Trad
Num. 3rd person Vs / Num. finite Vs	Morph
Avg. length of a T-unit	Syn
Avg. length of a Sentence	Syn/Trad
Complex Nominals per Clause	Syn
Complex Nominals per T-unit	Syn
Num. PPs per sentence	Syn
Avg. length of a clause	Syn

Table 8: The ten most predictive features according to Information Gain

Most of the features in the TOP 10 belong to the syntactic feature group, but morphology features and some traditional measures are included as well. The dominance of syntactic features, especially of those from SLA research, confirms the conclusions of Vajjala and Meurers (2012) for English that these measures are particularly effective for readability classification. The list also indicates that investigating the usefulness of morphological features for readability classification was well-worth the effort, even at this relatively shallow level of morphological modeling. Note also that the three traditional readability features (TRAD) that we view as a baseline are among the TOP10 features. While these superficial features have little conceptual value, they seem to be good predictors of reading level.

5.4 Results for Feature Groups and Combinations

We trained classifiers with various individual feature groups and some of their combinations. Table 9 summarizes the results for the different classification models. For the feature group combinations, only the most successful combinations are shown.

Feature set	Num. Features	Accuracy
TRAD	3	82.2%
LEX	23	82.1%
SYN	26	76.8%
MORPH	95	85.4%
LM	12	77.6%
SYN & MORPH	120	86.7%
LEX & LM & MORPH	130	89.4%
ALL	155	89.7%
TOP 10	10	84.3%

Table 9: A comparison of all five feature groups

Among the models trained on a single feature group, the morphological classifier (MORPH) performed best with an accuracy of 85.4%. The lexical classifier (LEX) (82.1%) and the baseline classifier (TRAD) (82.2 %) performed almost equally well, but slightly worse than the morphological classifier (MORPH). The classifiers trained on language model features (LM) and syntactic features (SYN) proved to be less effective predictors when taken on their own. However, they proved to be valuable when combined with other feature groups. Our experiments combining different feature groups showed that the syntactic (SYN) and morphological (MORPH) feature groups together were the most predictive two group combination with 86.7 % accuracy. When using three feature groups a combination of lexical (LEX), language model (LM) and morphological (MORPH) features performed best (89.4%).

Overall, the best result was achieved by combining all feature groups (ALL), which resulted in 89.7% accuracy. Compared to the traditional readability measures (TRAD) as baseline (82.2%), our best model improved classification performance by 7.5%. The classifier built with the ten best features (TOP 10) at 84.3% accuracy performed at about the level of the best single feature group (MORPH).

Conclusion and Outlook

As empirical basis of our work, we created the *GEO-GEOlino* corpus, a German corpus with two different reading levels that we collected from magazine articles that were available online. The *easy* reading level consists of the *GEOlino* articles targeted at children, while the *GEO* articles targeted at adults were labeled as *difficult*.

We trained classifiers with syntactic, lexical, and language model features derived from research on English, to see how well they can predict the reading level of German texts. We then introduced language-specific morphological complexity indicators as an additional group of features. We inspected a broad set of inflectional properties for German and for the first time made use of the derivational and inflectional morphology of nouns as features for readability classification of German. The novel morphological features proved to be especially good indicators for reading level, outperforming all other feature groups, when considered in isolation.

While all the individual feature groups except morphological features performed below the baseline, combinations of various feature sets resulted in higher accuracies. The best performance was obtained by combining all features. This achieved an accuracy of 89.7%, which is 7.5% above a baseline classifier using only traditional readability measures.

In terms of outlook, we are investigating how well the trained models generalize to other data sets, for which obtaining more graded reading material for German is an important next step. Going beyond readability classification of entire documents, we want to explore which features are effective not only at the document level but already at paragraph or sentence levels. Being able to identify simple and difficult paragraphs or sentences is relevant for identifying targets for simplification – the next step for our overall goal of building text simplification systems.

Finally, some of the features used in this paper were originally developed as measures of language proficiency in Second Language Acquisition research. Given how well SLA measures of language proficiency (based on texts produced by the learners) work as features for readability classification of native texts, it would be natural to take the features developed for our readability research back to the SLA domain and explore their applicability to classifying the language proficiency of language learners. In addition to quantitatively testing their impact for proficiency classification, strengthening that link could also help with qualitatively interpreting and further refining the different features on the background of SLA insights into stages of language development.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback on the paper, and Niels Ott and Ramon Ziai for the RFTagger Java interface. The research leading to the reported results has received funding through the European Commission's 7th Framework Program under grant agreement number 238405 (CLARA, `http://clara.uib.no`).

References

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California.

Bech, G. (1955). *Studien über das deutsche verbum infinitum*. Historisk-filologiske Meddelelser udgivet af Det Kongelige Danske Videnskabernes Selskab. Bind 35, no. 2, 1955; Bind 36, no. 6, 1957; Kopenhagen. Reprinted 1983, Tübingen: Max Niemeyer Verlag.

Bennöhr, J. (2005). A web-based personalised textfinder for language learners. Master's thesis, School of Informatics, University of Edinburgh.

Chall, J. S. and Dale, E. (1995). *Readability Revisted: The New Dale-Chall Readability Formula*. Brookline Books.

Chen, S. F. and Goodman, J. T. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–394.

Collins-Thompson, K. (2011). Improving information retrieval with reading level prediction. In *SIGIR 2011 Workshop on Enriching Information Retrieval*.

Collins-Thompson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

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

Dell'Orletta, F., Montemagni, S., and Venturi, G. (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*, pages 73–83.

dos Santos Marujo, L. C. (2009). Reap.pt (reap-portuguese). Master's thesis, Universidade Tecnica de Lisboa.

DuBay, W. H. (2006). *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Feng, L. (2010). *Automatic Readability Assessment*. PhD thesis, City University of New York (CUNY).

Fleischer, W. and Barz, I. (1995). *Worbildung der deutschen Gegenwartssprache*. Niemeyer, Tübingen, Germany.

François, T. and Fairon, C. (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*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-07)*, pages 460–467, Rochester, New York.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications at ACL-08*, Columbus, Ohio.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3.

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly*, 4(3):195–202.

Kincaid, J. P., Fishburne, R. P. J., Rogers, R. L., and Chissom, B. S. (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, Naval Technical Training Command, Millington, TN.

Klerke, S. and Søgaard, A. (2012). Danish parallel corpus for text simplification. In *In Proceedings of Language Resources and Evaluation Conference (LREC), 2012*.

Lau, T. P. (2006). Chinese readability analysis and its applications on the internet. Master's thesis, CUHK, Hongkong.

Levy, R. and Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

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

Lu, X. (2011). The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Languages Journal*.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, University of Memphis.

Meurers, W. D. (2000). *Lexical Generalizations in the Syntax of German Non-Finite Constructions*. Phil. dissertation, Eberhard-Karls-Universität Tübingen. Published as: Arbeitspapiere des SFB 340, Nr. 145.

Miltsakaki, E. (2009). Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, pages 49–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Napoles, C. and Dredze, M. (2010). Learning simple wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ott, N. and Meurers, D. (2010). Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications*, 3(1–2):9–30.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

Rafferty, A. N. and Manning, C. D. (2008). Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, PaGe '08, pages 40–46, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 777–784, Stroudsburg, PA. Association for Computational Linguistics.

Schwarm, S. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 523–530, Ann Arbor, Michigan.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, pages 574–576. ACM.

Skut, W., Kreen, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fith Conference on Applied Natural Language*, Washington, D.C.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904, Denver, USA.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In Tetreault, J., Burstein, J., and Leacock, C., editors, *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7) at NAACL-HLT*, pages 163—-173, Montréal, Canada. Association for Computational Linguistics.

Vor der Brück, T. and Hartrumpf, S. (2007). A semantically oriented readability checker for german. In *Proceedings of the 3rd Language & Technology Conference*.

Vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008a). A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429—-435.

Vor der Brück, T., Helbig, H., and Leveling, J. (2008b). The readability checker delite. Technical Report Technical Report 345-5/2008, Fakultät für Mathematik und Informatik, FernUniversität in Hagen.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.