Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment

Zarah Weiss, Xiaobin Chen, Detmar Meurers
Department of Linguistics & LEAD Graduate School
University of Tübingen
Germany
{zweiss,xchen,dm}@sfs.uni-tuebingen.de

Abstract

We investigate the readability classification of English and German reading materials for language learners based on a broad linguistic complexity feature set supporting the parallel analysis of both German and English. After illustrating the quality of the feature set by showing that it yields state-of-the-art classification performance for the established OneStopEnglish corpus (Vajjala and Lučić, 2018), we introduce the Spotlight corpus. This new data set contains graded reading materials produced by the same publisher for English and German, which supports an analysis comparing the linguistic characteristics of texts at different reading levels across languages. As far as we are aware, this is both the first readability corpus for German L2 learners, as well as the first corpus with comparably classified reading material for learners across multiple languages.

After discussing the first results for a readability classifier for German L2 learners, we show that the linguistic complexity analyses for the cross-language experiments identify features successfully characterizing the readability of texts for language learners across languages, as well as some language-specific characteristics of different reading levels.

1 Introduction

The language input available to language learners is a driving force for Second Language Acquisition (SLA), and reading is an important source of language input. Material that is just above the level of the learner is assumed to be best for fostering learning, which depending on the SLA tradition is characterized as i+1 input of Krashen (1981), input in the Zone of Proximal Development in socio-cultural approaches (Lantolf et al., 2015), or input reflecting second language development in usage-based SLA approaches (Ellis and Collins, 2009). Note that the focus here is not just on input that is understandable and of interest to the learner but also rich in developmentally proximal language properties.

This dependency of readability on reading purpose and individual language skills makes the identification of appropriate reading materials a major challenge for educators, especially for heterogeneous learning groups. Automatic readability assessment may facilitate the retrieval of appropriate reading materials for individual language learners. It refers to the task of identifying texts that are suitable for a given group of target readers with a specific reading purpose (Collins-Thompson, 2014). Recent approaches to automatic readability assessment also investigate the use of neural networks (Martinc et al., 2019). However, the identification of linguistic characteristics that impact the readability of texts in itself can also yield valuable insights for education, because it may inform content creators of reading materials for language learning. This also is an interesting research endeavor from a linguistic perspective and speaks against solely focusing on neural approaches. Similarly, it remains to be investigated to which extent these linguistic characteristics may generalize across languages given comparable target groups and reading purposes.

While there has been a considerable amount of work on automatic readability assessment for English, there is still insufficient research on other

Zarah Weiss, Xiaobin Chen and Detmar Meurers 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings 177: 38–54.

languages. The lack of suitable training corpora for other languages remains as one major limiting factor (Collins-Thompson, 2014), despite some research efforts to facilitate unsupervised readability assessments (Benzahra and François, 2019; Martinc et al., 2019). For example, there has been some recent work on German readability classifiers for native speakers (Weiss and Meurers, 2018; Weiss et al., 2018; Dittrich et al., 2019). Yet, a lack of corpus resources has so far hindered the development of a readability classifier for German as a second or foreign language (L2) learners.

In this article, we introduce a novel cross-lingual feature collection for broad linguistic modeling of German and English complexity. Although neural classification approaches have been strongly represented in readability assessment, our literature review (see Section 2) shows that their success has been very much limited on the benchmark data we use for this study and fallen behind the feature-based readability classification approaches which are also providing deeper linguistic insights while requiring less computational power.[1] However, while broad feature collections for language-specific complexity modeling have been proposed for English (Chen and Meurers, 2019) and German (Weiss and Meurers, 2018), they are not applicable across languages. This has so far hindered the cross-lingual study of similarities between characteristics of readability. We first validate our approach by applying it to an established readability corpus for English (Vajjala and Lučić, 2018), before using it to train two readability classifiers for labeling English and German L2 reading materials resulting in the first readability classifier of this kind for German. For this, we introduce a novel data set of English and German reading materials for beginning, intermediate, and advanced learners of English and German, the Spotlight corpus. We address the following research questions:

1. Can we train a successful readability classifier for German and for English using broad complexity modeling?

2. Can these classifiers generalize beyond their training language to cross-lingual contexts?

3. Which linguistic features are relevant for the distinction of reading levels and how do they differ between English and German?

The article is structured as follows. First, we discuss related work on readability assessment of English and German (Section 2). Then, we introduce the novel Spotlight data set (Section 3.1) as well as the OneStopEnglish corpus (Section 3.2) which we use as benchmark data set. We proceed to introduce our approach to automatic complexity assessment and the feature set (Section 4) we use throughout our machine learning experiments (Sections 5 and 6). Finally, we compare the informativeness of individual complexity features on Spotlight for the discrimination of reading levels (Section 7) before we come to the conclusion (Section 8) and outlook (Section 9).

2 Related Work

Automatic readability assessment has a long history dating back to the first readability formulas developed in the early 20th century, see DuBay (2006) for an overview. Traditional readability formulas employ few surface text characteristics such as text, sentence, and word length (Flesch, 1948; Dale and Chall, 1948). They are still widely used especially in non-linguistic studies on web accessibility (Esfahani et al., 2016; Grootens-Wiegers et al., 2015), in information retrieval systems (Miltsakaki and Troutt, 2007; Chinkina et al., 2016), and for confirming the compliance of reading materials with specific accessibility guidelines (Weiss et al., 2018; Yaneva et al., 2016), such as Easy-to-Read materials.[2]

Over the last two decades, there has been a shift towards computational readability classification approaches based on machine learning techniques employing feature engineering with Natural Language Processing (NLP) methods, see Collins-Thompson (2014) and Benjamin (2012) for an overview. Among others, linguistic complexity features from SLA research (Vajjala and Meurers, 2012), word frequency measures (Chen and Meurers, 2017), and features of text cohesion (Crossley et al., 2017) from Writing Quality Assessment research (Crossley, 2020) were shown to be valuable features for readabilty assessemnt.

While most readability research focuses on English (Collins-Thompson, 2014), to a lesser degree these approaches have also been employed for other languages such as Russian (Reynolds, 2016),

[1] See Strubell et al. (2019) for a discussion of the considerable energy demands of deep learning approaches in NLP.

[2] https://www.inclusion-europe.eu/easy-to-read/

French (François and Fairon, 2012), Swedish (Pilán et al., 2015), Italian (Dell'Orletta et al., 2013), or German (Vor der Brück and Hartrumpf, 2007). For German, the most recent classification approach has been proposed by Weiss and Meurers (2018) who use broad linguistic complexity modeling of German to distinguish between German media texts targeting adults and children. However, this approach only provides a rather coarse binary distinction and identifies reading materials for information retrieval (i.e., with a focus on accessibility), rather than language learning (i.e., with a focus on challenging the reader's language competence). Given the lack of appropriate multi-level reading corpora, so far no classifiers for German L2 readers have been trained.

Recently, several neural network approaches have been proposed for readability assessment (Martinc et al., 2019; Madrazo Azpiazu and Pera, 2019). Martinc et al. (2019) investigate the performance of supervised and unsupervised neural readability classification approaches for English and Slovenian. They find that their neural approaches perform overall at the state-of-the art level of feature-based classification approaches in both languages. For the OneStopEnglish corpus, their best classifier reaches an accuracy of 78.71% which performs at the same level as the feature-based classifier reported by Vajjala and Lučić (2018) with an accuracy of 78.12%. With this, the performance of neural approaches on OneStopEnglish does not exceed the original benchmark and lies substantially below the current state-of-the art on this data set, which is held by a feature-based classifier with an accuracy of 90.09% (Bengoetxea et al., 2020). In other words, while neural classification approaches have been very successful in several NLP tasks, they are currently not competitive with the breadth and depth of analyses supported by feature-based approaches to readability classification.

Only little research has been conducted on multilingual readability classification. While there are some neural classification approaches that are developed to be applicable across languages (Martinc et al., 2019; Madrazo Azpiazu and Pera, 2019), feature-based approaches are usually language-specific. An exception is the study by De Clercq and Hoste (2016), who compare the informativeness of lexical, semantic and syntactic features for English and Dutch readability classification. The cross-lingual applicability of multilingual models has so far not been investigated, except for a series of studies by Madrazo Azpiazu and Pera on the VikiWiki corpus, which distinguishes simplified Vikidia.org texts for 8 to 13 year old children from regular Wikipedia.org texts for Basque, Catalan, Dutch, English, French, Italian, and Spanish.[3] On this data, Madrazo Azpiazu and Pera (2020a) investigate the transferablility of the neural readability classification approach by Madrazo Azpiazu and Pera (2019). They demonstrate that training on multilingual data sets may improve readability classification results for low-resource languages in the binary classification task. Madrazo Azpiazu and Pera (2020b) follow a similar approach using a feature-based readability classification approach based on shallow features, morphological features, syntactic features, and semantic features. They report similar results as Madrazo Azpiazu and Pera (2020a). While these studies make an important first contribution to the assessment of cross-lingual readability assessment, they are clearly limited by the binary distinction of simplified texts for children and regular Wikipedia texts. The success of transfer learning for more fine-grained and practically relevant readability level distinctions remains to be empirically determined.

3 Data

3.1 Spotlight corpus

The Spotlight corpus consists of articles from the two monthly language learning magazines *Spotlight*[4] for adult German learners of English and *Deutsch perfekt*[5] for adult language learners of German. Both magazines are published by *Spotlight Verlag*, a leading European publisher for foreign language learning materials.[6] The magazines contain reading materials for beginning, intermediate, and advanced language learners which the publisher equates with the Common European Framework of Reference (CEFR) levels A2 (level: easy), B1/B2 (level: medium) and C1 (level: advanced).

We extracted all articles from the PDF versions of the respective issues provided to us for research purposes by the publisher.The type setting of the magazines made it impossible to di-

[3] https://github.com/ionmadrazo/VikiWiki
[4] https://www.spotlight-online.de
[5] https://www.deutsch-perfekt.com
[6] https://www.spotlight-verlag.de

rectly extract the individual articles with a PDF converter without loosing the information of their reading level. Instead, we manually identified and extracted each article using screenshots which we then converted to plain text using Google's optical character recognition (OCR) API.[7] This way, we extracted the English subset (henceforth Spotlight-EN) from the 110 issues of the *Spotlight* magazine that were published from January 2012 to December 2019 and the German subset (henceforth Spotlight-DE) from the 45 issues of the *Deutsch perfekt* magazine published from January 2018 to December 2019 (see corpus profiles in Table 1). The imbalance of readability levels in both data

Level	N. docs	N. sents	N. words
Spotlight-EN			
Easy	1.030	13.921	212.267
Medium	1.528	60.232	898.695
Advanced	1.030	24.288	440.793
$\sum$	3.285	98.441	1.551.755
Spotlight-DE			
Easy	763	16.135	180.178
Medium	509	27.107	338.553
Advanced	174	11.713	155.160
$\sum$	1.446	54.955	673.891

Table 1: Corpus profiles for Spotlight data

sets is due to the imbalanced distribution of reading levels in both magazines.

It is noteworthy that in both magazines, articles may vary considerably in length irrespective of their reading level. This is shown in Table 2. The table showcases that number of words – which has been and continues to be a popular surface feature for readability classification – is not sufficient to distinguish reading levels in this data set.

3.2 OneStopEnglish corpus

The OneStopEnglish (OSE) corpus by Vajjala and Lučić (2018) consists of overall 567 Guardian news paper articles that were rewritten for adult English as a Second Language learners by MacMillan Education.[8] Each Guardian article is available in an elementary (ele), intermediate (int), and advanced (adv) version resulting in a perfectly

[7]https://cloud.google.com/vision

[8]https://www.onestopenglish.com

	$\mu \pm SD$	M	Min	Max
Spotlight-EN				
Easy	206±166	137	53	877
Medium	588±555	493	23	4.497
Advanced	606±509	489	26	2.940
Spotlight-DE				
Easy	236±235	137	60	1.469
Medium	665±769	448	72	5.605
Advanced	892±537	524	91	4.161

Table 2: Article length in words in Spotlight data ($\mu \pm SD$= mean ± standard deviation; M = median; Min = minimal; Max = maximal)

balanced corpus.[9] The OSE corpus is a by now established reference data set for studies related to readability assessment and text simplification (Bengoetxea et al., 2020; Benzahra and François, 2019). Currently, the best results reported for OSE achieve an accuracy of 90.09% in a feature-based machine learning approach by Bengoetxea et al. (2020). Table 3 shows the corpus profile of the OSE data set. Table 4 displays the differences of article length across reading levels in OSE.[10]

Level	N. docs	N. sents	N. words
Ele.	189	6.033	105.169
Int.	189	6.634	128.335
Adv.	189	7.221	162.449
$\sum$	567	19.888	395.953

Table 3: Corpus profile for OSE

Level	$\mu(\pm SD)$	M	Min	Max
Ele.	556(±109)	561	267	948
Int.	679(±117)	691	315	1.083
Adv.	860(±171)	857	357	1.465

Table 4: Article length in words in OSE ($\mu \pm SD$= mean ± standard deviation; M = median; Min = minimal; Max = maximal)

[9]Since the three OneStopEnglish levels (elementary, intermediate, advanced) are not explicitly aligned with the CEFR levels, used to characterize the Spotlight levels (easy=A2, medium=B, advanced=C1), we keep the labels separate throughout the article.

[10]The numbers reported here slightly deviate from those reported by Vajjala and Lučić (2018), due to minor differences in the automatic tokenization.

As also noted by Vajjala and Lučić (2018, p. 299), there is a general tendency of articles becoming longer with increasing reading level. However, note the standard deviation of the article length within reading levels, which is considerable despite being much lower than the variability displayed in the Spotlight data.

4 Automatic Complexity Analysis

4.1 Complexity Features

We calculate 312 features of linguistic complexity merging the feature collections proposed by us in our previous work on German (Weiss and Meurers, 2018) and English (Chen, 2018). These have been successfully used for the tasks of readability assessment (Chen and Meurers, 2018; Weiss and Meurers, 2018; Kühberger et al., 2019), second language proficiency assessment (Weiss and Meurers, 2019b, 2021), academic language proficiency (Weiss and Meurers, 2019a), and teachers' grading objectivity (Weiss et al., 2019). While each of the feature collections contains more language-specific features than the joined feature collection proposed in this work, this is as far as we are aware the broadest collection of complexity features applicable to both, English and German, thus facilitating cross-lingual comparisons of complexity.

Our broad set of cross-lingual complexity features covers the theoretical linguistic domains of syntax, lexicon, and morphology, as well as features of discourse cohesion and psycho-linguistic features of human language use and human language processing. It also includes some surface measures from or inspired by classic readability formulas.

4.1.1 Surface Length (LEN)

We measure 21 surface text length features inspired by traditional readability formulas. They measure the raw number of sentences, syllables, letters, (unique) words including and excluding punctuation marks and numbers, and (unique) tokens. It also includes mean and standard deviations of sentence length and word length measured in letters, syllables, and words as well as the mean and standard deviation of words with more than two syllables. These categories can be applied without language-specific adjustments, except for the identification of syllables which are based on language-specific regular expressions.

4.1.2 Syntactic Complexity (SYN)

We assess several features of clausal and phrasal complexity that have been proposed in the SLA complexity literature (Wolfe-Quintero et al., 1998; Kyle, 2016) inspired by the implementations by Chen (2018) and Weiss and Meurers (2021). We measure 20 features of clausal elaborateness. This includes features measuring the length of clauses and (complex) t-units in various units (such as words, syllables, letters), as well as features of clausal coordination and subordination, such as the number of relative or dependent clauses per clause.

Furthermore, we measure 28 features of phrasal elaborateness. This includes several features focusing on the complexity of noun phrases (NPs) including the number of pre- and postnominal modifiers per complex NP, the number of (complex) NPs per clause, t-unit and sentence, and the length of NPs in words. It also entails features measuring the complexity of verb phrases (VPs) including the number of verb clusters and VPs per clause, t-unit and sentence and the length of verb clusters in words. We also measure the complexity of prepositional phrases (PPs) such as the number of (complex) PPs per clause, t-unit and sentence or the length of PPs in words. Finally, this includes measures of coordinate phrases per clause, t-unit and sentence.

While these syntactic features are identified based on language-specific TregEx (Levy and Andrew, 2006) patterns for constituency trees, we carefully designed all extraction rules to yield equivalent results across languages.

We also measure syntactic variation based on 12 measures of parse tree edit distances following Chen (2018).

4.1.3 Lexical Complexity (LEX)

We measure several complexity features assessing lexical richness, variation, and density that have been proposed in the SLA complexity literature (Wolfe-Quintero et al., 1998) inspired by the implementations by Chen (2018) and Weiss and Meurers (2021). These can be applied straight forward across languages as long as similar word categories (such as adjectives, nouns, verbs, etc.) can be identified.

This feature set includes 27 features of lexical density including POS-based lexical density features as well as 9 features of lexical diversity including lexical word, verb, noun, adjective, and

adverb variation. Finally, we assess 53 features of lexical richness including several mathematical transformations of type token ratios (TTR), parts-of-speech specific TTRs, the Uber index and HD-D (McCarthy and Jarvis, 2007).

4.1.4 Morphological Complexity (MOR)

Morphological complexity has been argued to be an important feature for readability assessment of morphological richer languages than English, such as German (Hancke et al., 2012; Weiss and Meurers, 2018) or Basque (Gonzalez-Dios et al., 2014). However, few measures have been used in readability assessment that are applicable across languages with different morphological systems. We use the Morphological Complexity Index (MCI) proposed by Brezina and Pallotti (2019) to assess morphological complexity independent of language by measuring the variability of morphological exponents of specific parts-of-speech within a text. These morphological exponents can be identified by contrasting word forms with their stems which makes the features applicable across languages. We assess overall 40 MCI features for verbs, nouns, and adjectives based on different number of samples and sampling sizes with and without repetition.

4.1.5 Discourse Cohesion (DIS)

We assess 26 features measuring the mean overlap of word forms and lemmas of lexical words, nouns, and grammatical arguments between sentences as well as their standard deviation. Each feature is calculated locally (between neighboring sentences) and globally (across all sentences in the text). These implicit cohesion features were originally proposed in CohMetrix (McNamara et al., 2014). Unlike explicit cohesion measures, such as the number of particular connectives, they are directly applicable across languages.

4.1.6 Language Use (USE)

Word frequency features have a long tradition in both, readability and complexity research. Yet, word frequencies obtained from different frequency data bases are not necessarily comparable. We address this issue by using the SUBTLEX-US (Brysbaert et al., 2011b) and SUBTLEX-DE (Brysbaert et al., 2011a) frequency data bases. We consider both SUBTLEX frequency data bases equivalent for the purposes of our complexity analysis because they represent word frequencies from the same register and were created to be maximally comparable. To mitigate effects due to the different sizes of the underlying corpora, we only use word frequencies per million words.

Based on this, we calculate 56 word frequency features including the mean (log) frequency of all words, lexical words, and function words and their standard deviations as well as frequencies for verbs, nouns, adjectives, and adverbs.

4.1.7 Human Language Processing (HLP)

Weiss and Meurers (2018) have proposed to use features based on theories explaining human sentence processing difficulties for readability assessment. They propose features based on the Dependency Locality Theory (Gibson, 2000) using the different integration cost weight configurations proposed in Shain et al. (2016). While the psycholinguistic theories have been formulated for English, the complexity features by Weiss and Meurers (2018) have so far not been applied for complexity modeling beyond German.

We implemented 21 features for both, English and German, based on universal dependencies to make them applicable across languages. These features calculate the average, maximal and highest adjacent discourse integration costs per finite verb across different weight configurations.

4.2 NLP Pipeline

We calculate our complexity features following a three-step procedure. First, we run a pipeline of Natural Language Processing (NLP) tools to provide linguistic annotations for the data. The annotation pipeline primarily relies on Stanford CoreNLP (Manning et al., 2014) which we use for sentence segmentation, tokenization, parts-of-speech (POS) tagging, constituency parsing, and dependency parsing for English and German. We additionally employ the Mate tools (Bohnet and Nivre, 2012) for lemmatization, because CoreNLP only provides a lemmatizer for English but not for German. We also use the OpenNLP Snowball stemmer to extract stems for English and German. For all annotations, we use the respective default models provided with the NLP tools.

Second, we count linguistic constructs using a set of extraction rules as well as word frequencies. This procedure is fully identical across languages except for syllable counts, POS-based counts, and syntactic complexity counts which we designed to be comparable across languages as described in

the previous section. For all other features we use identical extraction rules.

Third, we calculate a variety of complexity feature ratios based on these counts. This step is fully language independent.

4.3 Feature Extraction and Selection

We extracted all 312 features on OSE, Spotlight-EN and Spotlight-DE as described in the previous subsection. We then identified all features that were not variable on any of the three data sets. This way, we could exclude features that are irrelevant for the data sets while keeping the feature collections comparable across data sets. For this, we removed all features for which the most common feature value across all three data sets occurred in 95% of the data or more.

The feature removal reduced the entire feature collection to 301 features. Only human language processing features were removed through this step, including all features measuring high adjacent integration costs.

5 Establishing our Approach on OSE

5.1 Set-up

To validate the performance of our feature-based readability classification approach against an established benchmark data set, we first trained a classifier to predict reading levels on the OSE data. For this, we used the 301 complexity features from Section 4.3. All feature values were z-transformed and centered around zero. We trained a random forest (RF), an ordinal RF, a Support Vector Machine (SVM) with a radial kernel, and a SVM with a polynomial kernel in R (R Core Team, 2015) using the `caret` package (Kuhn, 2020).[11] In the following, we only report the results for the SVM using a polynomial kernel, which outperformed the other algorithms.[12]

To not reduce the relatively small data set further, we train and test using 10-folds cross-validation. We compare the performance of the classifier on OSE with a) the random accuracy baseline of 33.3% and b) the state-of-the art performance on this data set by Bengoetxea et al. (2020), reaching 90.09%. We also report the individual precision, recall and F1 scores for each reading level.

5.2 Results

The OSE classifier reaches an accuracy of 92.06% with a 95% confidence interval (CI) $= [89.52\%, 94.15\%]$ in 10-folds cross-validation. This significantly outperforms the random baseline of 33.33% (p-Value $< 2 \cdot 10^{-16}$).[13] It also exceeds the results of Bengoetxea et al. (2020).

Table 5 displays the confusion matrix for the classification summed across all 10-folds.

Pred\Obs.	Ele.	Int.	Adv.
Ele.	179	9	4
Int.	9	173	15
Adv.	1	7	170

Table 5: Confusion matrix: OSE 10-CV

It shows that misclassifications occur predominantly at adjacent reading levels and that there does not seem to be any systematic bias. Table 6 reports precision, recall, and F1 score per level. The performance across reading levels is relatively

	Ele.	Int.	Adv.
Precision	93.2	87.8	95.5
Recall	94.7	91.5	90.0
F1	94.0	89.6	92.6

Table 6: Performance for OSE 10-CV

balanced. Elementary texts have a slightly higher recall, while advanced texts have a higher precision. As expected when comparing an ordinal classification level with two adjacent levels with levels with only one adjacent level, intermediate texts receive the lowest scores for precision and recall.

6 Classifying Readability on Spotlight

6.1 Set-up

After establishing the performance of our approach against the OSE benchmark data set, we turn to our main research question, which compares feature-based readability classification across languages on Spotlight-EN for English and Spotlight-DE for German. Our classification is

[11] All R scripts, data tables, and trained models that are being reported in this and the following sections are publicly available on OSF at https://osf.io/5hbcs/

[12] SVM parameters: degree = 3, scale = 0.001, and C = 1.

[13] Here and throughout the article we report p-values obtained with one-sided t-tests with $H_1 = Acc. > Baseline$.

again based on the 301 complexity features we extracted and identified following the procedure described in Section 4.3. All feature values were z-transformed and centered around zero separately for Spotlight-EN and Spotlight-DE. This way, the classifiers are learning based on the standard deviations from the data sets' mean values rather than the raw feature values. This was supposed to mitigate language-specific differences, for example, regarding the average sentence length in German and English.

The set-up of the classification experiment is identical to the one described in Section 5.1. In the following, we only report the results for the ordinal RF which outperformed the other algorithms on both Spotlight data sets.[14] Since this is a novel data set, we use the majority baseline as sole reference to evaluate the classifier performance in the within language condition (Section 6.2.1).

For our cross-language classification experiment (Section 6.2.2), we apply the previously trained classifiers to the respective other subset of the Spotlight data, i.e., testing on Spotlight-DE for the classifier trained on Spotlight-EN and vice versa. Unlike previous cross-linguistic readability classification approaches that used cross-lingual data to augment limited training resources, this set-up tests the generalization of our classifiers in a form of zero-shot learning. We again compare the performance of each classifier across-languages against the majority baseline on the respective testing data and the within-language classification performance.

We also report the individual precision, recall and F1 scores for each reading level throughout all classification experiments.

6.2 Results

6.2.1 Within-language Performance

Table 7 displays the results of all four classification experiments on the Spotlight data. The Spotlight-EN classifier reaches an accuracy of 74.5% in 10-folds cross-validation. This significantly outperforms the majority baseline of 46.5% (p-Value $< 2.2 \cdot 10^{-16}$).

Looking at the confusion matrix in Table 8, we see that the classification is relatively balanced, even though in proportion to their total count advanced texts are classified incorrectly more often than the other reading levels. This can also be seen in the relatively low F1 score for advanced texts displayed in the first three rows of Table 10.

The Spotlight-DE classifier reaches an accuracy of 88.0% in 10-folds cross-validation. This significantly outperforms the majority baseline of 52.8% (p-Value $< 2.2 \cdot 10^{-16}$). Table 9 shows the confusion matrix for the classification, which shows good classification results throughout all reading levels. This is mirrored in the high precision and recall scores displayed in rows four to six in Table 10.

6.2.2 Cross-language Performance

For the classification across languages, the Spotlight-EN classifier reaches an accuracy of 55.5% on Spotlight-DE. Although this performance is considerably worse than for the within-language classification, this significantly outperforms the majority baseline of 52.8% (p-Value = 0.02118) showing that the classifier somewhat generalizes beyond English even if the performance drops considerably. Looking at the confusion matrix in Table 11, one of the most common misclassifications is the labeling of easy texts as medium. The classifier overestimates the reading difficulty of many easy and medium texts. This results in a high precision but low recall for easy texts, as shown in rows seven to nine in Table 10.

The Spotlight-DE classifier reaches an accuracy of 53.4% on Spotlight-EN. Again, this is much worse than the results for the within-language classification, but significantly outperforms the majority baseline of 46.51% (p-Value $= 1.284 \cdot 10^{-15}$). This shows again that the classifier generalizes to some degree in the zero-shot learning scenario. Looking at the confusion matrix in Table 12, it can be seen that the classifier tends to underestimate the reading difficulty of advanced texts (classifying them as medium or even easy) and of medium texts (classifying them as easy). This results in a relatively high recall for easy texts and very low recall for advanced texts, as shown in the final three rows in Table 10.

6.3 Discussion

The two readability classifiers trained on Spotlight-EN and Spotlight-DE are highly successful when applied within their training language and exceed the majority baseline con-

[14] Parameters for the English model: number of sets = 50, number of trees per div. = 150, number of final trees = 600; parameters for the German model: number of sets = 150, number of trees per div. = 150, number of final trees = 200.

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
Spotlight-EN	10-folds CV	74.5	$[73.0, 76.0]$	46.5	$< 2.2 \cdot 10^{-16}$
Spotlight-DE	10-folds CV	88.0	$[86.1, 89.6]$	52.8	$< 2.2 \cdot 10^{-16}$
Spotlight-EN	Spotlight-DE	55.5	$[52.9, 58.1]$	52.8	.02118
Spotlight-DE	Spotlight-EN	53.4	$[51.7, 55.1]$	46.5	$1.284 \cdot 10^{-15}$

Table 7: Overall classifier accuracy (Acc.) on Spotlight data compared against majority baseline (Maj.)

Pred\Obs.	Easy	Medium	Advanced
Easy	816	171	37
Medium	208	1,210	268
Advanced	6	147	422

Table 8: Confusion matrix Spotlight-EN 10-CV

Pred\Obs.	Easy	Medium	Advanced
Easy	727	83	1
Medium	34	399	27
Advanced	2	27	146

Table 9: Confusion matrix Spotlight-DE 10-CV

	Easy	Medium	Advanced
Spotlight-EN 10 CV			
Precision	79.7	71.8	73.4
Recall	79.2	79.2	58.1
F1.	79.5	75.3	65.0
Spotlight-DE 10 CV			
Precision	89.6	86.7	83.4
Recall	95.3	78.4	83.9
F1.	92.4	82.4	83.7
Spotlight-EN on Spotlight-DE			
Precision	82.3	42.5	52.4
Recall	44.6	67.4	67.8
F1.	57.8	52.1	59.2
Spotlight-DE on Spotlight-EN			
Precision	49.3	59.0	53.4
Recall	80.3	47.9	27.0
F1.	61.1	52.9	35.8

Table 10: Level-wise performance on Spotlight

siderably. When comparing the performance of the Spotlight-EN classifier and the OSE classifier, the different nature of the two English corpora has to be taken into account. OSE consists of the same 189 articles simplified for three different reading levels, which is a somewhat artificial set-up for training data. The Spotlight-EN corpus, instead, consists of different texts specifically written for a given reading level which is closer to real-life texts for which language learners might require automatic readability ratings. Thus, we consider the within-language performance of the Spotlight-EN classifier satisfactory.

Pred\Obs.	Easy	Medium	Advanced
Easy	341	73	0
Medium	408	343	56
Advanced	14	93	118

Table 11: Confusion matrix Spotlight-EN on Spotlight-DE

Pred\Obs.	Easy	Medium	Advanced
Easy	827	635	216
Medium	193	732	315
Advanced	10	161	196

Table 12: Confusion matrix Spotlight-DE on Spotlight-EN

For the Spotlight-DE classifier, we observe a very high performance throughout reading levels. Spotlight-DE is the first data set for the readability assessment of texts for German L2 learners that allows a distinction for beginning, intermediate, and advanced learners of German. Thus, we cannot compare the performance to a reference corpus or cross-corpus test the Spotlight-DE classifier. Overall, the classification results are sufficient to use the Spotlight-DE classifier in real-life scenarios, even though a cross-corpus evaluation on a comparable data set would be ideal to confirm its generalizability as soon as such a data set becomes available.

Turning to our cross-language classification experiments, we find that both classifiers generalize

to some extent in the zero-shot learning scenarios, despite considerable drops in performance. This result is not to be taken for granted due to the linguistic differences between English and German. These are highly promising initial results. Further research is needed to investigate to which extent this generalization also applies across other languages.

The comparison of the confusion matrices of both cross-lingual classification experiments reveals a symmetrical regularity in the misclassifications. While the German classifier underestimates the reading levels of the English texts, the English classifier tends to overestimate the readability of the German texts. Since the classifiers are trained and tested on feature z-scores centered around the mean this behavior is not immediately expected and warrants further investigation in future research.

7 Feature Informativeness on Spotlight

7.1 Set-up

To identify which of the 301 complexity features identified in Section 4.3 are most informative for the readability classification, we identify the most informative features using the correlation-based feature subset selection for machine learning approach by Hall (1999). This method identifies the subset of features that exhibits the highest correlation with the class to be predicted (in our case reading level) while minimizing the intercorrelation of features within the subset. We use the implementation provided in the WEKA toolkit version 3.9.5 (Hall et al., 2009) for feature identification. We report the percentage of features selected across each feature group before we discuss in more detail the intersection of features in both data sets.

7.2 Results

Table 13 displayed the raw number and percentage of features selected on Spotlight-EN and Spotlight-DE across feature groups and the total number of features contained in the feature group. To make the result summary more interpretable, we split syntactic and lexical complexity features into the individual subgroups distinguished within Sections 4.1.2 and 4.1.3. A full list of all features that are informative on either data set is displayed in Appendix A. Figure 1 shows the boxplots of all features that were selected for Spotlight-EN as well as Spotlight-DE.

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
LEX Density	7	(15.9)	5	(18.5)	27
LEX Diversity	1	(11.1)	1	(11.1)	9
LEX Richness	4	(7.5)	5	(9.4)	53
SYN Clausal	1	(5.0)	8	(40.0)	20
SYN Phrasal	1	(3.6)	5	(17.9)	28
SYN Variation	2	(16.7)	0	(0.0)	12
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11
$\sum$	49	(16.3)	43	(14.3)	301

Table 13: Informative features selected on Spotlight-EN (EN), Spotlight-DE (DE), and the total number of features in the group (All)

On Spotlight-EN and on Spotlight-DE, up to a third of all surface length features are selected, most of which are informative on both data sets. All of the shared length features increase with reading level (see Figure 1). Also language use features seem to be central for the distinction of reading levels on both data sets. 30.4% of the features were selected for Spotlight-EN and 19.6% for Spotlight-DE. Four of the language use features are relevant for both data sets: the average word frequency and its standard deviation are decreasing with increasing reading level. The same holds for the log frequency of lexical word types. The standard deviation of the verb token frequency is increasing with higher reading levels. Lexical complexity seems to play a medium role in the distinction of reading levels. 13.5% of the lexical complexity features were selected for Spotlight-EN and 12.4% for Spotlight-DE. Especially lexical density and richness play an important role on both data sets, but there is only very little overlap between the features selected for Spotlight-EN and Spotlight-DE. Only the POS density of modifiers and proper nouns as well as the squared word TTR were selected on both feature sets. For English, the proper noun density is decreasing, while the POS density for modifiers and the squared word TTR are increasing with reading levels. For German, the squared word TTR is also increasing with reading levels, but the two POS density features exhibit a u-shaped and inverse u-shaped

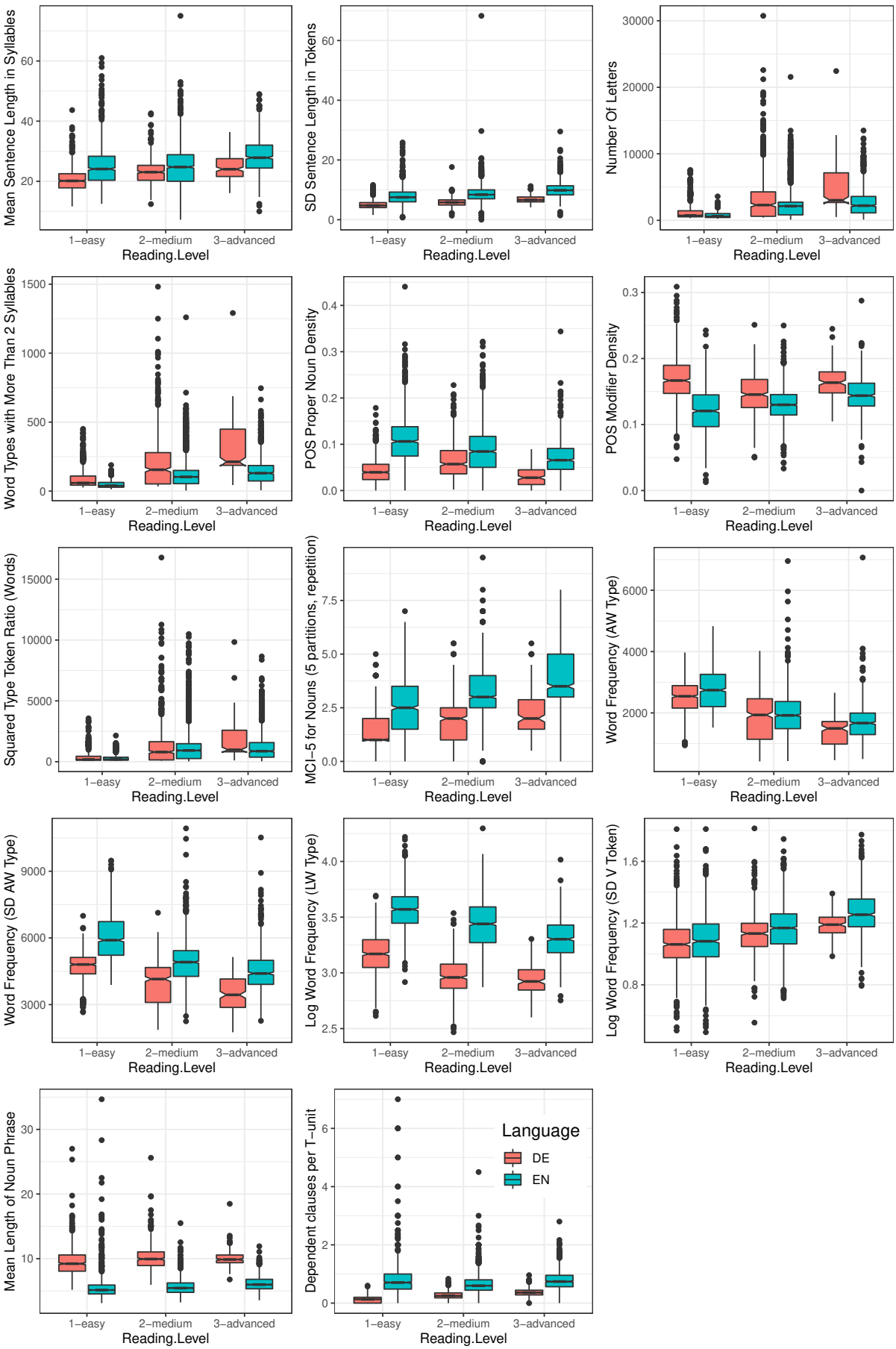

Figure 1: Boxplots of features that are informative on both, Spotlight-EN and Spotlight-DE

behavior.

The importance of syntactic and morphological complexity differs for Spotlight-EN and Spotlight-DE. Only 6.7% of the syntactic features were selected for Spotlight-EN, half of them features of syntactic variation. In contrast, 21.7% were selected on Spotlight-DE, all either features of clausal or phrasal complexity. Correspondingly, there is very little overlap in this domain between English an German. Only two syntactic features are informative for both data sets: the mean noun phrase length and the number of dependent clauses per t-unit, both of which are increasing with higher reading levels on both data sets. Morphological complexity features seem to play an important role for the distinction of reading levels on Spotlight-EN, but much less on Spotlight-DE. While 17.5% of the morphological complexity features were selected for Spotlight-EN, only 7.5% play a role on Spotight-DE. Both data sets share only one feature in this domain, namely the MCI for adjectives (measured with repetition with 5 partitions of size 5), which increases with higher reading levels, though the effect is more pronounced for English.

Neither implicit discourse cohesion features nor human language processing features seem to be important features on Spotlight-DE and also on Spotlight-EN, only 8.2% of the cohesion features were identified as informative.

7.3 Discussion

The correlation-based feature subset selection shows that features from most feature groups contribute meaningful information for the distinction of reading levels on both data sets. Especially features of surface length, language use, and lexical complexity help to characterize reading level differences on both data sets. Morphological and syntactic complexity features seem to capture more language-specific differences. There is also a considerable overlap of features selected for both data sets. Overall 28% of the features selected for Spotlight-EN and 32% of features selected for Spotlight-DE are shared between both data sets.

Judging from the features that are shared between the feature selections for English and German, higher reading levels are characterized by the use of less frequent vocabulary, longer words, sentences, and texts, and shifts in lexical density and richness. Also the features that were selected from the domains of morphological, phrasal and syntactic complexity increase with higher reading levels. This is in line with previous findings by Weiss and Meurers (2018) regarding the readability of German media texts targeting German-native speaking adults and children. However, our results indicate that these domains play a much less pronounced role for the distinction of reading levels. Interestingly, morphological elaboration seems to be more important for English than for German.

Human language processing measures do not seem to play an important role for the distinction of reading levels in either data sets, even though these measures are motivated by psycho-linguistic studies on human sentence processing. This is again in line with previous findings reported by Weiss and Meurers (2018).

Overall, these findings explain the albeit limited cross-language generalization of both readability classifiers in the zero-shot learning experiments. While there are differences in the types of features that are informative for the identification of reading levels across languages, there is nevertheless a substantial overlap and the shared features predominantly exhibit an increase in complexity with higher reading levels. This confirms that the publisher successfully instituted a policy facilitating the creation of stratified reading materials for different levels in a way that is comparable across the different languages that we analyzed.

8 Conclusion

We have investigated the use of language-independent broad linguistic complexity modeling for the multi-level readability classification of English and German reading materials for language learners. Our first study designed to benchmark the performance of our methods on the established OneStopEnglish yielded new state-of-the art results, clearly showcasing the value of broad linguistic modeling for readability assessment. Our study also shows that for certain tasks, broadly linguistically informed feature-based approaches are in fact not only competitive with neural approaches but exceeding their performance.

We then introduced a novel multi-level reading corpus for English and German on which we trained two readability classifiers that yield are highly successful within their respective training language. With this, we present the first multi-level readability classifier for German. This is highly relevant, because the much more com-

monly proposed binary classification approaches distinguishing simple and regular language are too limited to be of practical relevance for the retrieval of reading materials that are appropriate to foster foreign language learning.

We then demonstrated the generalizability of the German classifier for comparable English data and the English classifier for comparable German data. This is a novel contribution to cross-lingual readability research, not only because of the multi-level classification but also because of we propose a zero-shot cross-lingual readability classification approach unlike previous work focusing on augmenting low-resource training data. This is a central contribution to readability classification research, especially for languages other than English, given the lack of appropriate training materials for many languages.

In our final study, we compared the linguistic properties characterizing differences in reading levels in English and German. Our findings show that for both languages, texts systematically differ between reading levels in terms of the frequency and lexical complexity. Language-specific characteristics of reading levels can be found in the syntactic, discourse and morphological domains. The publisher thus successfully adapts the reading materials for different proficiency levels across a variety of linguistic domains in a systematic way. This is not a trivial insight, since previous work demonstrated that school book publishers do not always succeed in the linguistic adaptation of reading materials for different target groups (Berendes et al., 2018).

Our findings clearly demonstrate the value of feature-based classification approaches not only for the study of linguistic phenomena but also for readability classification. We demonstrate the feasibility of broad language-independent feature collections and their potential for zero-shot cross-lingual learning.

9 Outlook

As we saw in Table 7, cross-language zero-shot learning showed a promising result for training on Spotlight-DE and test on Spotlight-EN and the other way round. It is arguable that although different languages may complexify in different linguistic aspects, the general rule of more elaborate linguistic components and more varied expression usually resulting in higher complexity still applies. As a result, it is highly likely that zero-shot cross-language learning would also result in good performance, but detailed approaches need to be further designed and tested in future studies including more languages.

Another direction for future research is to see how the readability levels decided by the publisher match L2 learners' actual perception of the texts' difficulty. Although our models have yielded high accuracy, if the standards used to determine the levels of the texts do not actually match the learners' perceived difficulty, the predicted results are meaningless. Vajjala and Lučić (2019) offer an interesting data set that may potentially be used to answer this question.

Acknowledgements

We are grateful to the publisher Spotlight Verlag GmbH for making their publications available to us for research purposes.

References

Kepa Bengoetxea, Itziar González-Dios, and Amaia Aguirregoitia. 2020. AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural*, 64:61–68.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

Marc Benzahra and Yvon François. 2019. Measuring text readability with machine comprehension: a pilot study. In *Workshop on Building Educational Applications Using NLP*, pages 412–422.

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2018. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518–543.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second language research*, 35(1):99–119.

Tim Vor der Brück and Sven Hartrumpf. 2007. A semantically oriented readability checker for German. In *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznań, Poland. Wydawnictwo Poznańskie.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011a. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58:412–424.

Marc Brysbaert, Emmanuel Keuleers, and Boris New. 2011b. Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2(27).

Xiaobin Chen. 2018. *Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research*. Ph.D. thesis, Eberhard Karls Universität Tübingen Germany.

Xiaobin Chen and Detmar Meurers. 2017. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Xiaobin Chen and Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer-Assisted Language Learning*, 32(4):418–447. https://doi.org/10.1080/09588221.2018.1527358.

Maria Chinkina, Madeeswaran Kannan, and Detmar Meurers. 2016. Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 7–12, Berlin, Germany. Association for Computational Linguistics. http://anthology.aclweb.org/P16-4002.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of past, present, and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Scott A. Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1):11–28.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2013. Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose. In *Proceedings of Recent Advances in Natural Language Processing*.

Sabrina Dittrich, Zarah Weiss, Hannes Schröter, and Detmar Meurers. 2019. Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 41–56, Turku, Finland.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Nick Ellis and Laura Collins. 2009. Input and second language acquisition: The roles of frequency, form, and function. Introduction to the special issue. *The Modern Language Journal*, 93(3):329–335.

B. Janghorban Esfahani, A. Faron, K. S. Roth, P. P. Grimminger, and J. C. Luers. 2016. Systematic readability analysis of medical texts on websites of German university clinics for general and abdominal surgery. *Zentralblatt fur Chirurgie*, 141(6):639–644.

Rudolf Franz Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François and Cedrick Fairon. 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*. https://www.aclweb.org/anthology/D12-1043.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? Assessing the readability of Basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Petronella Grootens-Wiegers, Martine C. De Vries, Tessa E. Vossen, and Jos M. Van den Broek. 2015.

Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, 37(1):89–117.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Mark A Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbai, India. http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Stephen D. Krashen. 1981. The fundamental pedagogical principle in second language teaching. *Studia Linguistica*, 35(1–2):50–70.

Christoph Kühberger, Christoph Bramann, Zarah Weiss, and Detmar Meurers. 2019. Task complexity in history textbooks. a multidisciplinary case study on triangulation in history education research. *History Education International Research Journal (HEIRJ)*, 16(1). Special Issue on Mixed Methods and Triangulation in History Education Research.

Max Kuhn. 2020. caret: Classification and regression training. R package version 6.0-86.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

James P Lantolf, Stephen L Thorne, and Matthew E Poehner. 2015. Sociocultural theory and second language development. In *Theories in second language acquisition: An introduction*. Routledge New York, NY.

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

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020a. An analysis of transfer learning methods for multilingual readability assessment. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–100.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020b. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. http://aclweb.org/anthology/P/P14/P14-5010.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*.

Philip M. McCarthy and Scott Jarvis. 2007. A theoretical and empirical evaluation of vocd. *Language Testing*, 24:459–488.

Danielle A. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, Cambridge, M.A.

Eleni Miltsakaki and Audrey Troutt. 2007. Readx: Automatic evaluation of reading difficulty of web text. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*, pages 7280–7286, Quebec City, Canada. AACE. http://www.editlib.org/p/26932.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic assessment of language learning materials based on linguistic complexity. In *Proceedings of CICLING 2015- Research in Computing Science Journal Issue (to appear)*. https://arxiv.org/abs/1603.08868.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert Reynolds. 2016. *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications*. Ph.D. thesis, UiT - The Arctic University of Norway.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Ivana Lučić. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. http://aclweb.org/anthology/W12-2019.pdf.

Zarah Weiss, Sabrina Dittrich, and Detmar Meurers. 2018. A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Zarah Weiss and Detmar Meurers. 2019a. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019b. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Zarah Weiss and Detmar Meurers. 2021. Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, 7(1):84–131.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers. 2019. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Honolulu.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 49–57.

Appendix A: List of Selected Features

A.1: Features selected for Spotlight-EN

LEN Number Of Letters, SD Token Length in Letters, Percentage of Word Types with More Than 2 Syllables Length Measures, Number of Word Types with More Than 2 Syllables, SD Sentence Length in Tokens, SD Sentence Length in Syllables, Mean Sentence Length in Syllables

SYN Syntactic Complexity Feature: Dependent clauses per T-unit Clausal, Syntactic Complexity Feature: Mean Length of Noun Phrase Phrasal, SD Local Edit Distance for tokens, SD Global Edit Distance for Lemmas

LEX POS Density Feature: Particle, POS Density Feature: Adjective, POS Density Feature: Past Participle Verb, POS Density Feature: Article, POS Density Feature: Coordinating Conjunction, POS Density Feature: Modifier, POS Proper Noun Density, Corrected TTR, Corrected TTR Adjectives, Suqared TTR Words, Uber index (10) Adjectives, Lexical Verb Variation

MOR MCI-5 for Verbs (5 partitions no repetition), MCI-5 for Nouns (5 partitions no repetition), MCI-10 for Nouns (5 partitions no repetition), MCI-5 for Adjectives (2 partitions with repetition), MCI-5 for Adjectives (2 partitions no repetition), MCI-5 for Nouns (5 partitions with repetition), MCI-5 for Nouns (10 partitions no repetition)

DIS Global Lemma Overlap, Mean Local Noun Overlap (word form-based)

USE Logarithmic Word Frequency (Adj Type), Logarithmic Word Frequency (FW Type),

Logarithmic Word Frequency (SD Adj Token), Logarithmic Word Frequency (SD FW Type), Logarithmic Word Frequency (LW Type), Logarithmic Word Frequency (SD V Type), Logarithmic Word Frequency (AW Type), Word Frequency (AW Type), Logarithmic Word Frequency (V Type), Word Frequency (SD AW Token), Logarithmic Word Frequency (SD LW Token), Word Frequency (FW Token), Logarithmic Word Frequency (SD V Token), Logarithmic Word Frequency (Adv Token), Word Frequency (SD AW Type), Logarithmic Word Frequency (SD LW Type), Word Frequency (SD FW Type)

HLP *none*

A.2: Features selected for Spotlight-DE

LEN Number Of Letters, 2 Number of Word Types with More Than 2 Syllables, Mean Sentence Length in Syllables, SD Sentence Length in Tokens, SD Sentence Length in Letters

SYN Relative Clauses per Sentence, Relative Clauses per Clause, Dependent clauses per Sentence, Dependent clauses per T-unit, Complex T-unit Ratio, Dependent clause ratio, Relative Clauses per T-Unit, Mean Length of T-unit, Verb Cluster per T-Unit, Mean Length of Noun Phrase, Postnominal Modifier per Complex Noun Phrase, Verb Phrases per Clause, Verb Phrases per T-unit

LEX TTR Adverbs per Lexical Types, Squared TTR Nouns, Uber index (10) Verbs, Uber index (10) Nouns, Squared TTR Words, Modals per Verb, POS Modifier Density, POS To-infinitive Density, POS Possessive Pronoun Density, POS Proper Noun Density

MOR MCI-5 for Nouns (2 partitions with repetition), MCI-5 for Nouns (5 partitions with repetition), MCI-10 for Nouns (2 partitions no repetition)

DIS *none*

USE Word Frequency (V Type), Word Frequency (SD V Type), Logarithmic Word Frequency (Adj Token), Logarithmic Word Frequency (SD V Token), Word Frequency (AW Type), Logarithmic Word Frequency (SD Adv Token), Logarithmic Word Frequency (LW Type), Logarithmic Word Frequency (V Token), Word Frequency (SD FW Token), Logarithmic Word Frequency (SD AW Token), Word Frequency (SD AW Type)

HLP *none*