Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Automated Assessment of Language Proficiency on German Data

Edit Szügyi and Sören Etler and Andrew Beaton and Manfred Stede
Cognitive Systems Program
Applied Computational Linguistics
University of Potsdam / Germany
szuegyi|etler|beaton|stede@uni-potsdam.de

Abstract

The proficiency level of the learner is an important factor in various educational settings. In order to find the adequate language difficulty level, we classify texts written by language learners of German into proficiency levels A, B and C, as defined by the CEFR (Common European Framework of Reference for Languages), based on linguistic features extracted from the texts. Working on a combined data set of previously-used corpora, we use both data- and theory-driven feature sets, and determine the best-performing features. Our model achieves an accuracy of 82%, and the best-performing feature set contains features from all the theoretical groups, while all groups alone perform significantly above the random baseline.

1 Introduction

An important concept in the field of educational systems is Automatic Text Scoring (ATS), which automates the process of scoring texts by using NLP techniques. A special case of ATS is Automatic Proficiency Assessment (APA), which aims at scoring texts written by language learners according to a proficiency scale; in Europe, this is defined by the Common European Framework of Reference for Languages (CEFR). With the help of APA, educators can more easily find appropriate reading materials and students can get immediate feedback on their performance. Furthermore, perhaps we can also get closer to a more practical definition of the CEFR levels by way of linguistic feature extraction.

In the scope of this project, we have developed an APA system that classifies diverse German texts written by language learners into levels A, B and C of the CEFR. Level A (elementary), includes CEFR levels A1 and A2, Level B (intermediate), consists of levels B1 and B2 and level C (advanced) is composed of levels C1 and C2. We implement a wide range of linguistic features, which are described in Section 3.

2 Related Work

The earliest large-scale APA systems for German have been developed in the work of Hancke (2013) (see also Hancke, Vajjala and Meurers (2012)). She implements lexical, morphological, syntactic and language model features, building on work from different languages as well as different but highly related fields, such as Second Language Acquisition and Readability Assessment. Her feature sets are theoretically well-motivated and exhaustive. One aspect of her work that we think can be improved concerns the size and imbalance of the data set, the MERLIN (Wisniewski et al., 2011). While we also include it in our study, we overcome some of the problems by using a larger and balanced data set. Hancke achieved 72.5% accuracy working on 5 classes, A1–C1, and our overall goal is to build on and expand her research with new analyses.

As for other authors who work on German readability, Vajjala (2013) tests readability features on German text books in her PhD thesis, using the readability features developed by Hancke et al. (2012). Lavalley and Kay (2014) use children's writing as their data and work with embellishment clues (adjectives and adverbs) as features. Nietzio et al. (2012) work with texts written for mentally challenged people, and use sentence length and complexity features. Brück and Hartrumpf (2007) work with legal texts and semantic features. Zesch et al. (2015) use English and German texts to test which features are independent of the specific writing tasks or prompts. One very recent piece of work is by Weiss and Meurers (2018), who use media texts for children and adults with the goal of implementing a linguistically broad readability model for German. Their features are from the fields of lexicon, syntax, morphology, discourse, language use and human processing.

Among studies of texts written by learners of other languages, the majority – as can be expected – has focused on English. Yannakouadis (2011) works on grading texts written by learners of English with lexical features, part of speech (POS) tags, syntax features, length features and error rate. Treffers et al. (2018) study the correlation between lexical diversity and CEFR levels. Briscoe et al. (2010) analyze machine learning methods better suited for the task, using n-grams, parse rules, word length and error measures. It is also important to consider the possible end users of this line of research, namely educators, students, and readers. With that in mind, Chen and Meurers (2016) provide a publicly available platform for automatic complexity feature extraction and visualization.

3 Methods

3.1 Overview

In our work, we have implemented a wide range of features based on Hancke's thesis (2013), but using a bigger data set (see Section 4). While one of our goals is to develop a classification system, what we think is even more important is a thorough discussion of the performance of the feature space, and see how it relates to Hancke (2013), which is the only other piece of work discussing similar features in a similar setting.

We work with texts written by learners of German and implement a supervised classification model according to the CEFR categories A, B and C as labels. In order to train the model, we have experimented with different machine learning algorithms, such as Decision Trees, Logistic Regression and Support Vector Machines (SVM) with different configurations. We have decided to use a Linear SVM as it performed best given our data set. This was an expected outcome, as SVMs perform well in various classification settings, both inside and outside NLP. Other researchers in the field of automatic readability and proficiency assessment also found them to give the best results (Hancke, 2013; Pilán et al., 2016; Zesch et al., 2015; Weiss and Meurers, 2018)

We have used the implementation of scikit-learn (Pedregosa et al., 2011) with its default settings. Since SVMs are sensitive to the distribution of the data, we have balanced our data set. We end up working with 612 texts for each level, so our data set consists of 1836 texts altogether. When we present accuracy scores, they are based on a 10-fold cross-validation.

We use two different kinds of feature groups: data- and theory-driven. We have made this distinction because of the different approaches inherent in each. Namely, our theory-driven features are hand-engineered; we are checking for specific linguistic units or ratios we have theorized to predict proficiency. As for data-driven features, we are looking at the data as a whole, analyzing what we find by a given feature extraction method, without any concrete hypotheses. Our data-driven features are n-grams, parse rules, and grammatical tags, while our theory-driven features can be categorized into traditional, lexical, frequency, morphological, syntactic, and error measure sets. While this distinction is our own contribution, the specific features in the groups are mostly re-implementations of the features from Hancke's thesis (2013). The following is a general description of the sets, pointing out some important differences from her work.

3.2 Theory-driven Feature Sets

- **Traditional Features** predate advanced machine learning techniques. They are based on surface-level features, such as the average number of characters per word. While Hancke (2013) only works with text length, and the average number of words and syllables, we experiment with a wide range of traditional formulae, such as the Flesch reading-ease score or the SMOG score.

- **Lexical Features** measure the range and variety of vocabulary used in a text by a writer. A traditional measure is the type-token ratio (TTR). As the TTR is sensitive to text length, various mathematical corrections of the original formula have been proposed, such as the root TTR, corrected TTR, log TTR, Uber Index and Yules K. More recent attempts to account for this problem include for instance the hypergeometric distribution diversity (HD-D) (McCarthy and Jarvis, 2007) and the measure of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010). It is an interesting counterpoint to mention the work of Treffers-Daller et al. (2018), who claim that basic measures explain more variance in the CEFR levels of language learners' texts than the HD-D and MTLD, provided text length is kept constant across texts. Other lexical features are lexical density, measuring the ratio of lexical words to all words and lexical variation with respect to specific syntactic categories.

- **Frequency features** are based on the general idea that words that are more common in one language are acquired more easily and earlier by language learners. However, research also shows that especially L2 language learners often start with infrequent, more specific words (Crossley et al., 2011). We used a list with the number of occurrences of words in movie subtitles compiled by Brysbaert et al.(2011) to calculate the mean $log(frequency)$ and the standard deviation of the $log(frequency)$. The list was selected as subtitles are a common and easily available means of portraying everyday language. Hancke (2013) does not explain the choice of her binning method, so we chose equal width binning with 14 bins to determine whether words in certain frequency bands are characteristic of a specific level of text.

- **Morphological Features** are often realized through use of several linguistic features such as gender, case markers, verb tense markers, prefixes and suffixes. German is considered to be a morphologically rich language due to its three genders (masculine, feminine, neuter), four cases (nominative, genitive, dative, accusative), verb prefixes (both separable, such as *auf-*, and inseparable, such as *ver-*), and word compounding. In order to extract morphological features from the data set, we used the RFTagger (Schmid and Laws, 2008). For compound word detection, the CharSplit module for German was implemented (Tuggener, 2016).

- **Syntactic Features** measure the complexity of the dependency and parse tree structure of the text, based on Hancke (2013). She adapts these from various sources and fields and also adds some German-specific concepts to the feature set, such as the number of infinitival phrases with *zu*, or the ratio of passive constructions. For parse tree complexity, examples include the length of production units measured by average length of sentences and clauses, the number of clauses per sentence, or ratios of dependent clauses, coordinating conjunctions, and complex nominals per clause and sentence. In addition, we have also included the ratio of separated verb prefixes. As for dependency features, a representative feature is for instance the maximum and average number of words between a head and a dependent in a text.

- **Error Measures** As we are dealing with data written by learners of the language, we have implemented a spell check that also counts the number of misspelled and corrected words. We use this number to calculate the ratio of misspelled words and total number of words. Implicitly the error measures are part of some of our other feature groups as well, for instance the RFTagger uses the tag *FM* ('foreign word') for words it does not recognize as German, and in our application, many of those would actually be misspelled words.

3.3 Data-driven Feature Sets

- **Parse Rule Features** Following Hancke (2013), Briscoe et al. (2010), and Yannakoudakis et al. (2011), we build a feature vector out of Parse Rule frequencies for each text. An example would be 'NP ART NN' standing for a Noun Phrase consisting only of an article and a noun.

- **N-grams** are a theoretically simple yet powerful set of features to extract from unstructured data, which are used in a wide variety of NLP tasks, as the words used in a text intuitively convey a lot of information about its makeup. In the field of automatic test scoring, Yannakoudakis et al. (2011) and Briscoe et al. (2010) have worked with them. While unigrams are powerful, they are not capable of handling phrases, but it is easy to improve them by adding bi- and trigrams to the feature sphere. In this project, we implemented word, lemma, character and POS n-grams.

- **Grammatical tags** are extracted with the RFTagger. Some examples are the type of particles (answer, degree, negation, *zu*, separated verb particle) or the type of conjunctions (comparative, coordinating, subordinating with finite clause, subordinating with infinitive). Note that Hancke (2013) does not work with n-grams or grammatical tags.

4 Dataset

In order to overcome the limited availability of appropriate data, we use a combined data set built from five different sources: MERLIN (Wisniewski et al., 2011), Falko (Reznicek et al., 2012), KanDeL (Vyatkina, 2016), CLEG13 (Maden-Weinberger, 2013) and data from online sources for learners of German.[1] While all data sets contain different annotations, for the purposes of this project, only the CEFR level of the learner and the raw text were considered. As for the reading materials, we have decided to include them, as according to Pilán et al. (2016), textbook data can be beneficial for proficiency assessment in the event of a lack of data from the same domain.

The MERLIN corpus consists of texts written in an exam setting, which are assigned levels A1-C2 of the CEFR by trained human examiners. KanDeL is a collection of texts written by students from the US, who are enrolled in a basic German language program. The Falko corpus consists of text summaries written by C1-C2 learners of German and essays written by upper intermediate and advanced learners in various international institutions. Learners who scored more than 80 points on the C-test were hand-selected to form part of our C-level instances. The CLEG13 texts are essays, summaries and critical commentaries, and were written by students from the UK and labelled according to the year group of students into three groups. In the first two, students are assumed to be at levels B1 and B2, while the third consists of C1 learners.

See a summary of the combined data set in Table 1.

We are aware that the labels A, B and C do not necessarily signify the exact same level within the subcorpora. We have studied the levels' official description (Council of Europe, 2001) and the human-graded essays, and noticed that there can be significant differences inside one level. Thus, we believe that the categories are wide enough to allow for the potential differences caused by the non-uniform labeling methods.

5 Results and Discussion[2]

5.1 Theory-driven features

When calculating all our theory-based features, we arrive at a total of 129. See Fig. 1 for a PCA (Principal Component Analysis) graph of the data set. PCA is a method for dimensionality reduction of data by retaining as much variance (information) as possible. It is easy to note that while A and C are neatly separated, level B is more interspersed throughout the graph. This result is intuitively plausible: While it is easier to give a casual definition for a beginner or an advanced speaker, the intermediate level, by its very definition, is a less well-defined category in between the two.

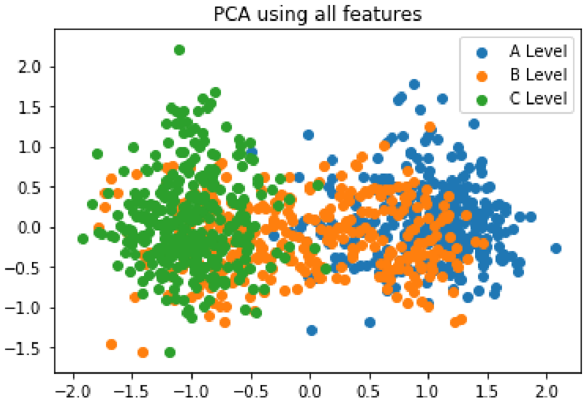

Figure 1: PCA with our theory-driven features

Our results are not directly comparable to any of the literature we have encountered, mostly due to the different class labels used. Hancke (2013), our main background literature, classifies according to A1-C1 on the MERLIN data set. We have re-implemented a large part of her best performing features, which she calls Best34 (Hancke, 2013, p. 56). She achieves 72.5% accuracy and we achieve 69% in the same setting. Our assumption is that the difference is due to the different spell checker we use, as the MERLIN data is very noisy compared to our other data sources. Hancke (2013) uses Google Spell Check[3], which was not publicly

[1] german.net/reading/, lingua.com/german/reading/, www.cornelsen.de/shop/capiadapter/download/get/

[2] One can argue that there is a conceptual overlap between the theory- and data-driven feature sets. However, the two feature sets are kept distinct throughout the experiments so it should not affect the results.

[3] https://code.google.com/p/google-api-spelling-java/

	Level A	Level B	Level C
CLEG13		416 (172,876)	315(146,545)
FalkoSummaries			107 (40,787)
FalkoEssay			159 (84,519)
FalkoWHIG			92 (56,513)
KanDeL	185(29,635)		
MERLIN	363(22,576)	624 (103,986)	
Reading	64 (10,390)	41 (8,642)	
Total	**612(62,601)**	**1,081(285,504)**	**673(328,364)**

Table 1: Number of texts (and tokens) in the subcorpora

Feature set	**Data Set**	**Classes**	**Acc.**
Best34	MERLIN	A1-C1	**72.5**
Best34 by us	MERLIN	A1-C1	69
Our best	MERLIN	A1-C1	70
Our best	Combined	A-C	**82**

Table 2: Results

available at the time of our project. Using our best features on the same data in the same setting, we achieve an accuracy of 70%. For a comparison, see a list of our best features in Table 3 and Best34 in Hancke (2013, p. 56). Testing on our newly created dataset, the model achieves a 82% accuracy, and most importantly, it very rarely misclassifies A for C, or vice versa. The verification of this statement is shown in the confusion matrix in Figure 2. The figure also shows that the model makes the highest number of incorrect predictions when classifying level B. See a more detailed error analysis in section 5.3, and a summary of the results in Table 2.

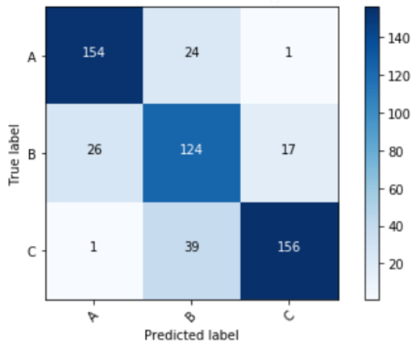

Figure 2: Confusion matrix presenting the classification results on a test set.

As for the analysis of our feature groups, when tested alone, traditional features, lexical features, morphological features and syntactic features achieve an accuracy above 70%, with morphology being the best group. The reason for that might be that German is a morphologically complex language, and also that some of our morphology features capture the complexity of other fields as well. For instance the different verb forms, while morphologically different, also represent syntactic complexity. Even the spelling error features alone achieve an accuracy that is significantly better than the random baseline of 0.33.

When performing an iterative feature elimination, we arrive at around 40 features that perform approximately equally as well as our whole feature set. Additional features do not increase the classification accuracy significantly. The 40 best consist of 5 traditional, 8 lexical, 2 frequency, 5 spelling error, 14 morphological and 6 syntactic features.

These features are similar to Hancke's(2013) best-performing group, with the main difference being that she does not work with the traditional readability features, and we are not using language model features. See the features that performed best in this project in Table 3.

It is interesting to note that all of our spelling error features are in the best group, signaling that analysis of the errors the writer makes is a good direction for future additions to the feature set. While working with the data, we have noticed that the frequency of spelling errors noticeably changes across corpora. Ideally, the setting in which the text was written should also be taken into account, as the MERLIN texts which contain the most spelling errors were written in an exam setting, while other corpora also include homework assignments.

Inside the syntax group, we can see that the general complexity features perform well, e.g., average number of words between the head and a dependent, average clause length, number of

Group	Features
Traditional	SMOG, average number of characters per word, number of polysyllables, FOG
Lexical	**Lexical Diversity** Yule’s K, Uber Index, HDD, MTLD
	Lexical Density and Variation adverb variation, modifier variation, verb variation, corrected verb variation
Frequency	bin 0, bin 6, mean frequency
Error	spelling errors, spelling errors with correction, capitalization errors, umlaut spelling errors, real spelling errors
Morphological	number of articles, ratio of compound nouns to noun, number of 1st person tags, number of past tense tags, ratio of nominative to nouns, number of nominatives, ratio of *-keit* suffix to nouns, number of past participle verbs, ratio of participles to verbs, number of singular tags, ratio of dative nouns to nouns, number of second person tags, ratio of verbs per sentence, ratio of 1st person to finite verbs
Syntactic	**Dependency** average number of words between head and dependent, average number of dependents per noun excluding modifiers
	Parse Tree Complexity average clause length, average number of dependent clauses per clause, average number of non-terminals per sentence, average number of interrogative clauses per sentence

Table 3: Best-performing features

non-terminals, or the ratio of dependent clauses per clause. From the more specific features we can see that NP complexity is relevant.

As for morphology, we can see that both compounds and derivational features (nouns ending with the suffix *-keit*), as well as inflections appear in the best-performing features. Certain verb forms, like past tense, participles, 1st and 2nd person have a correlation with proficiency. The inflection of nouns is also relevant, and the number of datives and nominatives divides the data best.

One can see that a lot of the features are dependent on sentence length, e.g. the traditional readability measures SMOG and FOG, number of polysyllables or most of the lexical features. The degree to which sentence length influences the classification level is up for debate. While it is true that it is theoretically possible to produce a short but complex or a long but simple text, in real life scenarios text length and complexity or proficiency very often go hand in hand. Exploring this correlation further is a direction for further research.

5.2 Data-driven features

We have experimented with word, lemma, character and POS n-grams. For words and lemmas, instead of a simple count, a tf-idf (term frequency-inverse document frequency) weighting method is used, with the goal of scaling down the effect of features that occur very frequently and are thus not informative, for instance *the*. We set the length of n-grams to 3 to avoid the problem of data sparsity. We have iteratively experimented with the number of features that gives the best accuracy.

According to, for instance, the findings of Zesch et al. (2015), when analyzing the task-dependency of features, n-grams are highly task-dependent. Thus, as expected, our n-gram model performs rather poorly when trained on one subdataset and tested on another. See Fig. 3 to observe how closely n-grams are related to the data set and specific tasks that the learners were writing about: Words such as *Kansas* show up since one corpus is written by students from Kansas, as well as *Feminismus* (’feminism’), as that was one of the essay topics. Some features are more generalizable, for instance *ich* (’I’) is an important feature for A level text, which relates to the communicative skills needed for beginner levels, i.e., they are expected to be able to talk about their immediate surroundings. In fact, all of the words support this observation. Inside the negative coefficient group, we can see the complementiser *dass* (’that’), showing a syntactic feature; low-level learners are likely to not use dependent clauses.

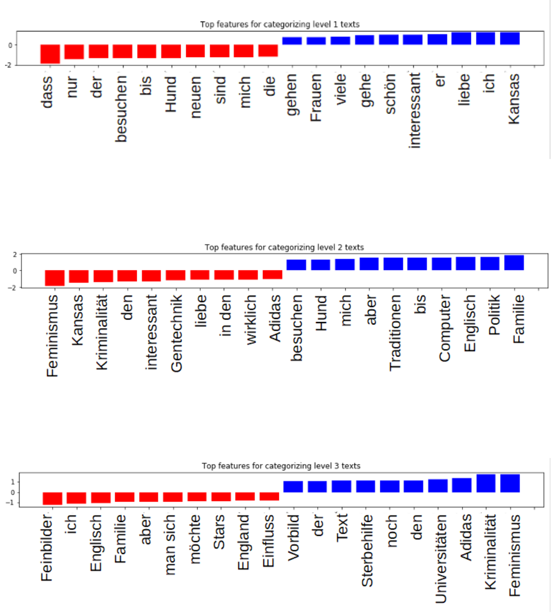

Figure 3: Important word n-grams

POS-unigrams, while producing a low, 59% accuracy, do present some interesting observations. For level A, some of the most predictive features are the presence of FM (foreign material), probably due to misspellings, VFIN (finite verbs), or INT (interjections). The VINF (infinitival verb) shows up in the features for levels B and C, as they are part of more complex verb structures, and the use of a PROADV (pronominal adverb) or VPP (participle verb) signals a level C.

See Table 4 for the accuracies of our different n-gram features. Trained on a specific task, they could achieve really high accuracy and we notice interesting observations looking at their results, however, they generalize poorly. Note that the cross-dataset accuracy is binary.

N-gram	#Feat.	Accuracy	Cross-data
Word	10,000	0.756	0.66
Lemma	5,000	0.748	0.63
Char.	20,000	0.881	0.55
POS	5,000	0.835	0.65

Table 4: Accuracy of n-gram features

The parse rules (PR) are extracted by the Stanford Parser. In order to reduce computational complexity and increase relevance, we have excluded rules that appear fewer than 10 times in the data set. With this we arrived at 1222 parse rules. The length of the parse rules can be anything greater than or equal to two. We have achieved the best accuracy with 200 PR features, which was 0.766 +/- 0.056. See Table 5 for some of the best-performing PR features and possible interpretations. The best accuracy we have reached is 77% with 500 PR features. We notice that the data-driven features support and validate our theoretical feature engineering. NP and PP complexity seems to be important in classification, as is the use of conjunctions and *zu*-infinitives. The table is intended as an illustration of possible interpretation of some rules with high importance. For an exact understanding of the PR features, the TIGER Treebank (Smith, 2003) can be consulted.

Interpretation	Parse rules
NP complexity	NP PIAT NN NP ADJA NN NP ART NN NP ART NN
PP complexity	PP APPR ART ADJA NN PP APPR NN
Conjunctions	CNP NN KON NN CAP ADJA KON ADJ
Adj. and adv.	AVP ADV ADV AP ADV ADJD
Zu-Infinitive	VZ PTKZU VVINF

Table 5: Important Parse Rules and their interpretation

Prediction using the 85 grammatical tags of the RFTagger gave an accuracy of 79 (+/-4) %. The tags name, masculine, full verb, noun and coordinating conjunction are the best predictors for level A; attributive adjectives, personal pronouns, prepositions, and degree particles signal B level texts the strongest, while C level texts are best recognizable by colons, interrogatives, adverbs, negations and definite articles, according to the model. The presence of the word *zu* (in English corresponding to 'for', 'to' or the intensifier 'too') is the clearest sign of a text not being A level. We can conclude that in the case of this data-driven feature set as well, many of the features the system found to be important are the same as those we manually created. Some additional features, such as different kinds of articles, pronouns, or particles can be added to the model for future experimentation.

5.3 Error analysis

In order to think about directions for improving our classifier in the future, we have performed an error analysis on incorrectly classified sentences.

As we are dealing with a 3-way classification, the biggest error the classifier can make is a miss of two levels, e.g. classifying A instead of C. Running a classification 10 times and observing and analyzing these errors, we can state that the number is very low, ranging between 0 and 3 at each trial. An example sentence from one of these texts is *Dieser Film handelt die amerikanische revolutionäre Zeiten während der frühen Monate Jahres 1776 , die auch ihm seine Name gibt.* (This film is about the American Revolutionary Times during the early months of 1776, which also gives it its name.) The text is part of the KanDeL data set, which includes both in-class assignments and homework. Our assumption is that the text was written as a homework assignment, so the writer could put time and effort into performing beyond their expected proficiency level. When observing the feature values for the text compared to the mean values for A texts, not only does it surpass them in the surface-level categories, but also in categories such as average number of words between head and dependent. Another example of misclassification is when there is not enough information in the text, for instance the A level text *LIEBER JENS, GLÜCKWUNSCH* ('DEAR JENS, CONGRATULATIONS') gets classified as C by our classifier. Due to its length, the relatively long term *Glückwunsch* or the high TTR are possibly given too much weight as many of the other features would be zero. Another text that showed up multiple times in this non-exhaustive experiment was a C-level text from the CLEG data that was classified as A. On closer inspection, comparing the feature values to the mean, the problem is the surface-level features, like number of syllables or SMOG. The implications of these findings is that the traditional readability formulas are not without their problems, which we are aware of. However, excluding them completely from the calculation is not the best option, as they show up in the most distinctive features.

The question of errors by one level is more complicated. We can conclude that texts from the CLEG data set are mostly classified higher than their label. This might be unintuitive as in the case of CLEG, the labels are actually the level the students are supposed to be at a certain academic year, and not dependent on any test score or essay grade. Working with the data, we have already noticed that CLEG level B is closer to the level C from other sources. Texts from the MERLIN data are the most commonly misclassified in both directions. We assume that level A texts often get misclassified because of the lack of information they contain. We also have to note that our classifier at the moment does not take the flow or cohesion of texts, or correct word choice, into account, which would probably be vital when dealing with such a short text; looking at the feature vector, we see that such a text tricks our classifier in terms of surface-level and lexical features which are highly correlated with text length.

6 Conclusion

With the help of our classifier and the data set, the writing level of language learners can be found with a reasonably high accuracy. We found that linguistic features correlate with CEFR proficiency levels and can perform reasonably well in a classification scenario. Moreover, with our detailed description of the performance of different features, we hope to have come closer to a tool that helps educators obtain a more practical list of what is expected from learners at certain levels. In the case of German, our target language, morphological features appear to be especially important. Some syntactic and lexical features are also given a high weight by the machine learning algorithm.

By constructing a larger and more balanced data set, we report 82% accuracy, a significant improvement over our models performance on just the MERLIN texts, which reached 70%. For further investigation, the most important factor is the data set itself. With enough data, we can also try running the model on the full CEFR scale from A1-C2, instead of the three-level classification currently being performed. Additional improvements to the data preprocessing can also be incorporated into our current pipeline, such as experimenting with different German spell checkers and sentence boundary detection methods. As for the feature groups, data from other fields, such as semantic or pragmatic information, are not included in the scope of this project; this additional information would also be worth testing. In other cases, a more fine-grained feature division could be helpful, for instance, analyzing different error categories.

It is also important to note that the feature sets, while tailored for German, are not language-specific per se. The data-driven features are all language-neutral and as for the theory-driven ones, traditional, lexical, frequency, and error measures are also not tied to the language of the text. As German is a morphologically rich language, it is unsurprising that morphological features perform well for classification, which may not be the case for other, morphologically less rich languages. As for syntactic features, most features are related to the dependency or parse tree structure and thus, also language neutral. However, a few features, such as the number of passive constructions, or specific infinitival phases we would not expect to contribute to the results greatly in other languages. Testing cross-language performance is a promising direction for future research, as well.

References

T. Briscoe, B. Medlock, and O. Andersen. 2010. Automated assessment of ESOL free text examinations. Technical report, University of Cambridge Computer Laboratory.

M. Brysbaert, M. Buchmeier, M. Conrad, A. Jacobs, J. Bölte, and A. Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in german. *Experimental Psychology*, 58:412–424.

X. Chen and D. Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. in proc. of the workshop on computational linguistics for linguistic complexity.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

S. A. Crossley, T. Salsbury, D. S. McNamara, and S. Jarvis. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4):561–580.

J. Hancke, S. Vajjala, and D. Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proc. of COLING 2012: Technical Papers*, pages 1063–1080.

J. Hancke. 2013. Automatic prediction of CEFR proficiency levels based on linguistic features of learner language. Master's thesis, University of Tübingen.

R. Lavalley and K. Berkling. 2014. Data exploration of sentence structures and embellishments in german texts: Comparing childrens writing vs literature. In *Proc. of the 12th edition of the KONVENS conference, Hildesheim*, volume 1.

U. Maden-Weinberger. 2013. CLEG13 corpus of learner german. documentation.

P. McCarthy and S. Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24:459–488, October.

P. McCarthy and S. Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

A. Nietzio, B. Scheer, and C. Bühler. 2012. How long is a short sentence? - a linguistic approach to validation and definition of rules for easy-to-read material. In K. Miesenberger, A. Karshmer, J. Klaus, and W. Zagler, editors, *Proc. of Computers Helping People with Special Needs, 13th International Conference, ICCHP 2012*, pages 369–376.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

I. Pilán, E. Volodina, and T. Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *COLING*.

M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K. Schmidt, H. Hirschmann, and T. Andreas, 2012. *DasFalko-Handbuch. Korpusaufbau und Annotationen. Version 2.01.*

H. Schmid and F. Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. *COLING 2008, Manchester, Great Britain.*

G. D. Smith. 2003. A brief introduction to the tiger treebank, version 1.

J. Treffers-Daller, P. Parslow, and S. Williams. 2018. Back to basics: How measures of lexical diversity can help discriminate between cefr levels. *Applied Linguistics*, 39(3):302–327, 04.

D. Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zurich, Faculty of Arts.

S. Vajjala. 2013. *Analyzing text complexity and text simplification: connecting linguistics, processing and educational applications*. Ph.D. thesis, University of Tübingen.

T. vor der Brück and S. Hartrumpf. 2007. A semantically oriented readability checker for german. In *Proc. of the 3rd Language & Technology Conference, Poznan, Poland*, page 270274, October.

N. Vyatkina. 2016. KANDEL: A developmental corpus of learner german. *International Journal of Learner Corpus Research*, 2(1):102–120.

Z. Weiss and D. Meurers. 2018. Modeling the readability of german targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proc. of the 27th International Conference on Computational Linguistics (Coling 2018)*.

K. Wisniewski, A. Abel, and D. Meurers. 2011. Merlin. multilingual platform for the European reference levels: Interlanguage exploration in context.

H. Yannakoudakis, T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *HLT '11*, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

T. Zesch, M. Wojatzki, and D. Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proc. of the Building Educational Applications Workshop at NAACL*.