Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 707–714
Marseille, 20-25 June 2022

Subjective Text Complexity Assessment for German

Laura Seiffe[1], Fares Kallel[1], Babak Naderi[2], Sebastian Möller[1,2], Roland Roller[1]
[1]German Research Center for Artificial Intelligence (DFKI),
Speech and Language Technology Lab, Berlin, Germany
firstname.secondname@dfki.de
[2]Quality and Usability Lab, TU Berlin, Berlin, Germany
firstname.secondname@tu-berlin.de

Abstract

For different reasons, text can be difficult to read and understand for many people, especially if the text's language is too complex. In order to provide suitable text for the target audience, it is necessary to measure its complexity. In this paper we describe subjective experiments to assess the readability of German text. We compile a new corpus of sentences provided by a German IT service provider. The sentences are annotated with the subjective complexity ratings by two groups of participants, namely experts and non-experts for that text domain. We then extract an extensive set of linguistically motivated features that are supposedly interacting with complexity perception. We show that a linear regression model with a subset of these features can be a very good predictor of text complexity.

Keywords: Corpus Creation, Readability Assessment, Complexity

1. Introduction

The ability to read and to understand text is a crucial competence to communicate and to exchange information. For different reasons, understanding written language can be a challenge not only for different target groups like language learners, students, or people with cognitive limitations, but also for people with less experience or knowledge about the text's content. All these user groups have different needs in terms of readable texts. To enable people to understand text is not only important for general communication, but also plays a long-term role for inclusion and integration. Additionally, user-friendly and user-centring applications need text that is well adapted to the needs of their target groups. Besides understanding which texts are hard to read for specific target groups, the assessment of readability is an important intermediate target for text simplification - knowledge about the obstacles the different reader groups have, is crucial to make text easier. A machine based evaluation of text readability can be used to get insights into this quality aspect of text.

The definition of the term *readability* as well as the closely related terms *understandability* and *(text) complexity* are widely discussed in the field of research (Zamanian and Heydari, 2012). Linguists, educationalists, psychologists, and machine learning engineers have different perspectives on that. Whether the text is human made or machine generated plays a role for the definition of readability as well (Howcroft et al., 2020). Generally, readability is often very vaguely defined as how easily a written text is to read. In other cases, readability is equated with linguistic complexity. Understandability usually refers to the understanding of the text and examines the reader's perspective: The understanding of a text depends on their prior knowledge of topic and language (Mclaughlin, 1969), (vor der Brück and Hartrumpf, 2007).

In the following, we use the definition of *readability* as an umbrella term that covers the concepts of *complexity* and *understandability*, as suggested in Chall and Dale (1995), DuBay (2007), and Naderi et al. (2019). Thus, complexity is the specification of linguistic features that influence the understanding of a text and understandability is the dimension of understanding the text's message, considering the reader's background knowledge. This means that the reader's perspective and the linguistic characteristics of a text cannot be easily separated - and we assume that there is no need to do so. Evaluating the linguistic characteristics of a text with respect to their influence on the reader's understanding enables user-centered results.

We make two hypotheses for this paper. First, we expect that the target group has a significant influence on the complexity rating. We expect that non-experts will rate test sentences as more difficult than experts. Second, we expect complexity ratings to correlate with measurable and extractable linguistic features. This means that a complexity score per sentence can be automatically derived based on a combination of these features. If the first hypothesis is correct, it will follow that there must be an individual complexity score per target group. Thus, the complexity assessing might depend on different linguistic feature combinations.

In this paper we present a corpus of German sentences, which has been annotated with readability assessment scores. The annotation was carried out within two experiments, by two different target groups, "experts" and "non-experts". A statistical difference in the rating behaviour between the target groups was found. In order to examine the prediction of readability assessment scores, we explore a wide range of lexical, syntactic, and morphological features, and test them with a baseline model on our data. The results show that a linear regression model with a set of 20 features is able to predict the readability of users to a good accuracy. The corpus, as well as the developed features will be published with this work. To the best of our knowledge, this is the first corpus in German, addressing subjective readability ratings of a German target group other than

German learners.

2. Related Work

Research on text readability has a long history (DuBay, 2006). The interest increased with the rise of readability formulas in the 1970s - early studies such as the Flesch-Kincaid readability formula (Kincaid et al., 1975) use superficial linguistic features, such as word and sentence length. Those studies were designed for English text and addressed students as target groups, using the readability level of school grades as readability classification.

In the last two decades, readability assessment research benefited from the increase of available data and the possibility of extracting more sophisticated textual features. Especially the combination of a broad and profound set of linguistic features and machine learning applications advanced the outcome of readability prediction. For English, diverse sets of linguistic features as well as text corpora were compiled and examined to predict text complexity, taking aspects of educational sciences, psycho-linguistics, and second language (L2) acquisition research into account (see, e.g., Schwarm and Ostendorf (2005), Lu (2010), Feng et al. (2010), Vajjala and Meurers (2012)).

Linguistic features that interact with complexity are highly language dependent. While so far most research has been done for English, there is important progress in other languages as well. For example, Broda et al. (2014) presented experiments on Polish text. Chatzipanagiotidis et al. (2021) conducted a readability classification experiment on Greek textbooks, Nassiri et al. (2018) presented an approach for the readability assessment of Modern Standard Arabic and Imperial and Ong (2020) worked on Filipino storybooks. Battisti et al. (2020) compiled a corpus consisting of parallel as well as monolingual, simplified German texts for the use of automatic readability assessment and automatic text simplification.

Most research focuses on general purpose texts or merges different text types into one corpus, e.g., the WeeBit corpus (Vajjala and Meurers, 2012) or its baseline dataset, the WeeklyReader corpus (Schwarm and Ostendorf, 2005). For more specialized, technical language, Loughran and McDonald (2014) did experiments on financial text, and Severance and Cohen (2015) examined the readability of medical abstracts in order to specifically research the complexity of technical language.

The readability of a text depends not only on its language related features but also to a large extent on the individual perception of the reader. Taking this subjective perspective of the reader into account, the concepts of *readability* and *understandability* blend. Often, either students or L2 learners are targeted. This follows the tradition of the readability formulas and seems understandable: First, there is available data, either in form of school box texts for different grades (Mulyanti and Soeharto, 2020), (Petersen and Ostendorf, 2009) or in form of parallel corpora for different age groups (Hancke et al., 2012), (Gala et al., 2020). Second, as language learners are a clearly defined target group, user studies can be conducted relatively successfully (Naderi et al., 2019). Third, there is undoubtedly a need for assessing readability for these target groups. Alternatively, the perspective of an unspecified, "general" audience is taken into account. The classification of the text complexity depends then usually on expert ratings of linguists and/or language professionals (Kate et al. (2010) or De Clercq et al. (2014)). The definition of a more specific target group and the focus on a specific text domain mean that the readability classification gets a subjective tendency. However, it is to question whether complexity assessment generally can be objective, because the reader, the text domain and the complexity rating interact with each other (Dale and Chall (1949), DuBay (2004)). Addressing this subjectivity seems useful for developing user-centred applications that are supposed to work in a predefined context. It also means that the classification of readability should not (only) depend on inherent text properties or evaluation of professionals, but should also adapt the perspective of the targeted reader group as accurately as possible, as discussed by Vajjala and Lucic (2019). This is achieved if the target group itself contributes to the complexity classification of the training data. In this paper we present a German corpus of domain specific sentences that are annotated with complexity ratings of two target groups. We further present experiments for predicting this subjective readability perception on the basis of an extensive set of linguistically motivated features.

3. Corpus Creation

In the following we describe our text data, as well as the two different readability rating experiments that were conducted to label our corpus.

3.1. Source Text

Our dataset consists of text provided by our project partner, a German IT service provider in the context of German tax consultants, auditors, and lawyers. This text is splitted into sentences and originates from instructions, commentaries and descriptions which address employees of the service provider, as well as external users of the system. They describe technical solutions to the company's products or give more detailed descriptions about law regulations affecting the company's clients. There is no limitation in accessing the texts for outsiders.

The creation of the texts underlies strict quality rules, also in terms of complexity. The authors have access to guidelines which, for example, provide a set of verbs that should be used in certain conditions, or explicitly state that certain grammar structures are to avoid, e.g., passive constructions or the nominalization of verbs.

We used 232 documents and splitted them into sentences. We choose the sentence level to explore the complexity perception on that level specifically; alternatively, paragraph level or even document level would have been of interest. However, the paragraphs in the documents are rather small and interrupted by tables, pictures, enumerations, key words, and bullet points. By examining paragraphs or the whole document the level of text understanding would become more important. We assume that the sentence level makes it easier for the test person to rate the complexity without considering the actual context.

The extracted sentences are manually curated by a linguist expert to exclude text that is not understandable in isolated

form (without context), e.g., table or picture descriptions. Also, sentences that contain co-referential structures are excluded, in the cases where they are not understandable without context. After this cleaning process, the corpus contains 2929 sentences. Some examples are given in Table 1.

3.2. Subjective Readability Ratings

We conducted two experiments to annotate the corpus with subjective readability ratings - one by non-experts using crowdsourcing, and one by expert staff of the IT service provider. We hypothesize that the perception of text complexity depends to a large extent on the expertise of the reader, e.g., knowledge of the text domain.

In both experiments, we used the same survey design and interface to keep the experiments comparable. The rating is made on a 7-point Likert scale, going from *very easy* (1) to *very complex* (7) as suggested by Naderi et al. (2019), see Table 2.

We followed the Absolute Category Rating test methodology which is widely used in the domain of Quality of Experience for subjective assessment (Möller and Raake, 2014). We calculated the average of ratings provided by all participants per each sentence. The resulting metric is called Mean Opinion Score (MOS) and is a subjective measure of readability of each sentence.

3.2.1. Non-Expert Target Group

We conducted the experiment with non-domain-experts using the crowdsourcing platform Clickworkers[1], a platform with a sufficiently large group of German speaking crowd workers. Crowdsourcing provides access to a large scale of geographically distributed group of workers who can take part in a subjective test (Naderi, 2018). In comparison with laboratory based experiments, crowdsourcing is a faster, cheaper and more scalable approach (Naderi, 2018). Meanwhile, previous works show that crowdsourcing provides competitive results with laboratory tests, e.g., in Iskender et al. (2020).

To create a representative group of non-expert users, the crowd workers had to fulfill the following three conditions before answering the questionnaire:

1. German native or at least B2 level
2. Not a tax or law professional, and not an employee of the data providing German IT service provider
3. Passing a language test to assure the level of language knowledge

The language test was a German listening comprehension; three short audio files were played and the participants had to answer questions. Only participants who passed 90% at least were accepted. By these means, we tried to ensure that only participants with a profound knowledge in German took part. According to the self declaration, 98.9% of the non-expert participants were German native speakers, 0.9% were level C speakers and 0.1% were level B speakers. The passed language test seems to verify these numbers.

[1] https://www.clickworker.de/

Postulating more conditions would make it more difficult to find a satisfying number of crowd workers or simply reduce the amount of ratings. Also, a verification of the participants' self-declaration is hardly possible.

The remuneration per survey (10 sentences) was 0.8€.

To ensure the quality of the ratings, an intensive data cleaning was necessary (c.f. Table 3). We excluded raters who 1) did not pass the language test 2) obviously cheated by always providing the same rating, 3) rated a trap question higher than intermediate complexity and 4) showed a low inter-rater reliability, which means that they were not consistent when rating the same sentence. Aside from that, we excluded single ratings that had a low correlation with the mean of all ratings of that sentence. All in all, 61.94% collected ratings, from the non-expert target group, passed the data cleaning process.

3.2.2. Expert Target Group

The same experiment was conducted with employees of the data providing IT service provider. The hypothesis is that the experts (employees) are more familiar with the texts than the non-expert crowd, thus they should rate the complexity as overall easier.

The participants did not get any compensation, except a raffle of three vouchers among all participants. The employees were not asked to pass a language test in order to ease the process for them. Since the experts were the employees of our project partner, we could already assume a high level of German native speakers before the experiments took place. In our data set, 96.9% of expert participants are German native speakers (3.1% C level speakers).

Each expert rated at least 10 sentences; some few rated up to 120 sentences. However, for only 47 sentences, we got four or more ratings. See Table 3 for a corpus summary of both target groups.

3.3. Analysis of the Dataset

The percentage distribution of the rating classes and items (see Figure 1) shows that for both groups, experts and non-experts, there is a strong tendency towards "very easy" and "easy" ratings. Overall, the mean of all items is for experts 3.1 and for non-experts 2.7. Surprisingly, especially the non-experts rate more than 25% of items as "very" easy, although we assumed that the lack of domain knowledge would make the texts more complex to them.

Considering the two target groups *experts* and *non-experts*, our hypothesis is that the experts should perceive the texts significantly easier than the non-experts (null hypothesis) as they are reading the texts on a daily basis, and some of them are even the texts' authors.

The first step to test this assumption is to align the datasets. As there are fewer ratings of the experts (896) than of the non-experts (5.990), a sufficient comparability is not given for all sentences. We consider 47 sentences which are at least rated four times (mean = 4.3) by experts[2] and 12 times (mean = 16) by non-expert crowd workers.

[2] Although that is a small number of ratings for calculating MOS values, we still consider that because participants were volunteer, trustworthy experts.

Original Sentence	Translation
Alle Lohnarten, die bei den Mitarbeitern abgerechnet werden und sozialversicherungsrechtlich als laufender Bezug behandelt werden, müssen im Soll-Entgelt berücksichtigt werden.	*All wage types, which are settled in the employees and are treated as a relevant to social insurance law, must be taken into account in the target fee.*
Klären Sie dies im Vorfeld mit der zuständigen Behörde.	*Clarify this in advance with the responsible authority.*
Wenn die Mitarbeiter korrekt zugeordnet sind, besteht für Sie kein Handlungsbedarf.	*If the employees are correctly assigned, there is no need for action.*

Table 1: Some Example Sentences of the Final Dataset

How do you rate the overall complexity of this sentence?	
Very difficult	7
Difficult	6
Somewhat difficult	5
Neutral	4
Somewhat easy	3
Easy	2
Very easy	1

Table 2: Likert Scale for Complexity Rating

	Experts	**Non-Experts**
# participants	118	262
# ratings	1 970	9 670
# kept sentences after cleaning	371	370
# kept ratings after cleaning	896	5 990
Native speakers	96.9%	98.9%
Avg. number of ratings per item	2.8	17.3

Table 3: Overview about Complexity Rating Annotation

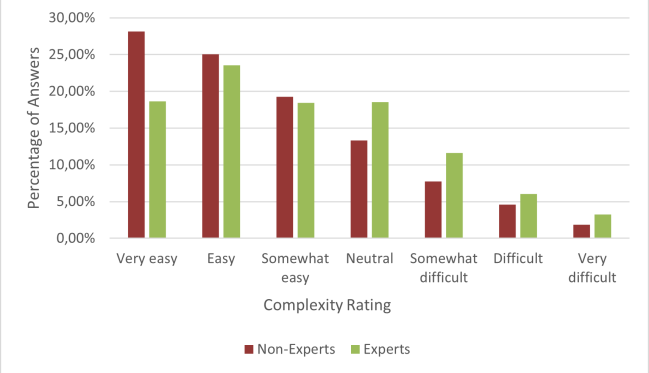

Figure 1: Percentage Distribution of Answers per Target Group

We fitted a linear mixed-effects model (LMEM) with random intercept and with expertise, and sentences as fixed factors and participants as random factors. The results show significant main effects of sentence $F(46, 533.55) = 19.961$ $p < 0.001$ and expertise $F(1, 90.26) = 4.201$ $p = 0.043$ with no interaction effect. Figure 2 illustrates the distribution of the ratings between expert and non-expert groups. Experts tend to rate the sentences to be slightly harder than non-experts. Although we expected a difference between the rating behaviour, we assumed the direction would be the other way round. The reason (or reasons) for this result cannot be fully explained at this point; only speculations are possible. One possibility could be that the experts are more concerned with the outcome of the experiment, i.e., the quality of their rating is higher. Another aspect could be a varying understanding of complexity: As the non-experts are lacking the context, they might just have assessed how complex the grammatical structure is. In contrast to that, the experts with domain knowledge could have rather answered how well they understood the sentence. Language proficiency is rather not a reason for the different rating, since the percentage of native speakers is almost identical for both groups.

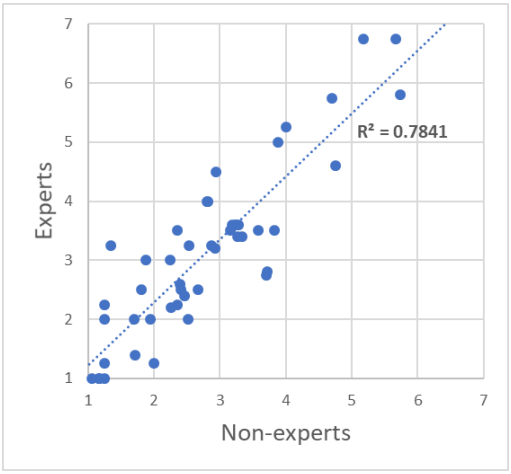

Figure 2: Distribution of ratings between expert and non-experts.

The results indicate that the assessment of readability does somewhat depend on the reader. However this conclusion should be considered with caution as the number of samples from experts was limited in this study and led to a significant but borderline effect. Because of that borderline result and the generally small number of expert ratings, we decided to merge the datasets for the next experiments. As we hypothesized that non-experts would rate the sentences as more complex, we must reject our first hypothesis. For future work, two adaptations of the setup might be considered: First, an extension of the dataset would be of interest. Second, a more fine grained selection of subjects might be helpful.

4. Predicting Subjective Text Complexity

Based on this dataset, we perform first experiments with the intention to predict the subjective text complexity rating. We want to find out whether a combination of various linguistic and easy to compute properties of the sentences can be used to predict the subjective assessment of the text complexity. In other words, we research whether there is a correlation between linguistic features and the subjective text complexity assessment.

In the following, we present a set of linguistically motivated features that are then used in first baseline experiments for modelling the prediction as a regression task.

4.1. Linguistic Features

Following the definition of readability we gave in the introduction, complexity reflects the level of readability on the linguistic level. Assuming that certain properties of language correlate with the assessment of readability, we extract various different features.

The features we consider in the readability prediction are mainly derived from the three linguistic levels, lexicon, morphology, and syntax. Besides, we also include features from earlier readability formulas (e.g., Kincaid et al. (1975)). As we are only examining text on the sentence level, discourse related features are neglected. In order to create a feature set that relates to German text complexity, we used the described features of Schwarm and Ostendorf (2005), Pitler and Nenkova (2008), Feng (2010), Lu (2010), Hancke et al. (2012) and Vajjala and Meurers (2012) starting point. Especially Hancke et al. (2012) provide a carefully curated set of features adapted to the German language. Overall, we compiled a list of 147 features which is provided along with the corpus and the script to extract those features. In the following, some selected features are described.

Lexical features describe properties on the semantic level. They refer to the vocabulary of the language. By the means of the lexical features we want to measure how semantically rich the sentence is, e.g., by calculating a type/token ratio or the information density. Also of interest is how common the used words are in the German standard language. To calculate this, we compare the words with the DeReWo list of German word forms (DeReWo, 2013) - a list that assigns frequency scores to German words. The calculated score indicates how frequent the word is in the general language - the more frequent it is, the easier it should be perceived. Next to a frequency score, lexical density can also be measured by the distribution of content words and function words. Function words are word classes that only serve grammatical needs, e.g. auxiliary verbs or determiners.

Morphological features seem especially interesting in German as the language has a rich inflectional and derivational morphology. We calculate various ratios concerning the inflectional behaviour of verbs. Following Hancke et al. (2012), we also count a set of suffixes in German that refer to nominalization processes - the assumption is that the nominalization of a verb (e.g. *nutzen* "to use" vs. *Nutzung* "usage") is perceived as more complex than the corresponding verbal structure. Furthermore, compounds are a very productive derivational process in German. They reflect a dense informational content: Within one word a whole relational meaning can be conveyed. We count compounds per sentence, and count the elements per compound.

On the *syntactic level*, we examine the complexity of the sentence structure. The syntactic features include the number of clauses and the average length of a clause per sentence. Again, we follow Hancke et al. (2012) in their differentiation of clauses and T-Units: A clause is a phrase with one finite verb and its corresponding subject, while a T-Unit is a clause (finite verb plus subject) plus a dependent, subordinate clause. Corresponding to the examination of clauses and T-Units, we also calculate the dependency depth. The more clauses a sentence contains, the higher is the assumed complexity.

Parts of Speech (PoS) are labels on the lexical-grammar interface. We calculate various ratios to express the PoS structure of each sentence in numbers. The distribution of specific PoS gives implicit insights into both the sentence's syntactic structure and its lexical density.

Features from the early *readability formulas* are examining the surface level of texts. They include number of tokens, number of characters and number of syllables per sentence. Overall, these features correspond to text length.

4.2. Feature Extraction

We developed a python based pipeline to extract all linguistic features for every sentence in the dataset. All features are calculated and stored as numeric tabular data for further analysis. For that purpose we mainly make use of the natural language processing library Spacy (Honnibal and Montani, 2017) which provides a number of pre-trained pipelines.The labelling of German word classes in Spacy relies on the Stuttgart-Tübingen Tagset (STTS, Schiller et al. (1999)). Using these tags, we were able to count the occurrences of different parts of speech and use that to extract meaningful features.

By making use of the syntactic dependency parser, we can show and navigate the grammatical structure of sentences and therefore build syntactic parse trees and define patterns to extract different syntactic features. For example, to extract the number or the length of clauses per sentence, we identify first the finite verb, second we identify all tokens that belong to the branch of that finite verb and third we reconstruct these tokens into the corresponding clause.

Spacy's pipeline also includes a morphologizer which holds the morphological characteristics specific to every language. This is useful for instance to extract the case of nouns, which in the German language can differ from accusative, dative, nominative and genitive and also analyse the inflectional behaviour of verbs. For compound splitting and counting, we use the python module "german_compound_splitter"[3].

Some of the calculated variables were normalized by the number of tokens in the sentence. This was done in order to maintain the linguistic relevance of the occurred phenomenon and avoid the influence of the sentence length.

[3] german_compound_splitter, Copyright 2020 by repodiac, see `https://github.com/repodiac/german_compound_splitter` for updates and further information

4.3. Preliminary Experiments

In the following, we examine if we can predict readability using a simple baseline model, in combination with the previously extracted features. Our task has a linear nature as the complexity of sentences increases with increasing values of the input features, which calls for a linear solution. Thus, we set the problem as a regression problem and attempt to predict the exact value of complexity using a Linear Regression model. We evaluate the model using the Root Mean Squared Error (RMSE) as a metric. RMSE is a commonly used measure for regression and it estimates the deviation of the predicted values from the expected values. In simple words, this means the lower the value is, the better is the performance.

To get the best results, the input feature set had to go through some preprocessing before being fed to the model. This preprocessing included the downscaling of large features to values between zero and one, and the removal of features with high autocorrelation values, as well as features which always had the same value across the complete dataset. Furthermore, as already discussed in Section 3.3 and shown in Figure 1, both rating groups tend to rate texts as easy. To avoid a bias, we used an equally distributed dataset (according to complexity rating) for the following experiments.

For our experiment, we are interested in finding out which feature classes or which single features hold the most relevant information when predicting the complexity. Therefore, we test different combinations of features based on the linguistic classes syntax, lexicon, and morphology. In addition, we also test the features based on the readability formulas.

Feature Set	# Features	RMSE
Morphological	28	0.49
Lexical	57	0.43
Syntactic	14	0.41
Readability formulas (RF)	5	0.35
Syntactic + RF	19	0.31
Selected features	20	0.20

Table 4: Results on our dataset using different feature combination sets, evaluated with Root Mean Squared Error (RMSE)

As a next step we attempt to find a subset of features that can produce the best accuracy in predicting the readability. To do so, we perform an ablation study on the features. In this way, we were able to reduce the number to 20 features belonging to different linguistic classes. Table 4 summarizes the results of all tested feature combinations after training and testing our model. When only taking homogeneous linguistic classes into account, no class (syntax, lexicon, or morphology) can outperform the results of the readability formulas features which reach an already satisfying error of 0.35. This value is slightly improved to 0.31 when readability formula features are combined with the syntactic features which perform better than morphological and lexical features. Fitting the selected 20 features to the model produced the lowest error value of 0.2.

4.3.1. Results and Discussion

Figure 3 shows the final result of the ablation study and indicates for each feature in this subset the relevance it has in the mapping function for predicting the complexity of a sentence. The most dominant feature is the number of words. This feature which simply relates to text length seems to play the most important role in text complexity. This correlates with the success of the isolated readability formulas features. Similar applies to the number of characters (*char_average*). Information density, expressed by e.g., the features function words (*function_words*), compounds per sentence (*compounds_per_sent*), noun type-token ratio (*Noun_TTR*) and adjective type token-ratio (*ADJ_TTR*), plays also a major role. The more new words a sentence has and the more information the words convey, the more complex the sentence is. Verbal morphological features seem to have a small to no influence, as the only relevant feature is the ratio of infinite verbs to all verbs in a sentence (*infinitive_to_V*). The syntactic complexity is represented by the features expressing the length of a verbal phrase (*Av_len_VP*), a nominal phrase (*Av_len_NP*) and a prepositional phrase (*Av_len_PP*) which might correlate with general text length - a "pure" syntactic feature with a rather large influence is the dependency depth (respectively height of the parse tree, *Abs_tree_height*).

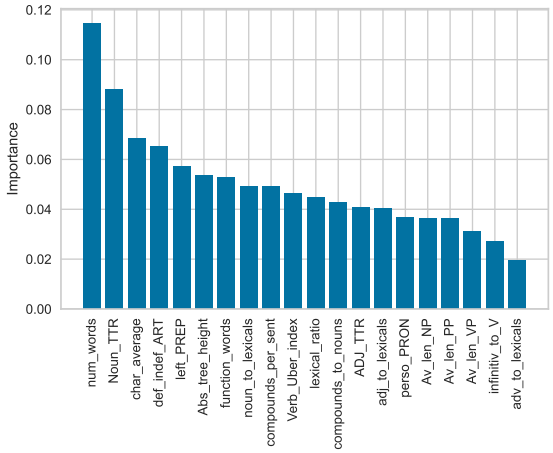

Figure 3: Feature Importance

5. Conclusion

In this work we created a dataset of German sentences that were annotated with subjective readability ratings by two different target groups - experts and non-experts. Results showed that the participant's expertise has a significant main effect on their ratings. Unexpectedly, experts tend to rate sentences to be more complex than the non-expert group.

As baseline experiments, we explored the influence of various linguistic features on this rating. We compiled a set of linguistically motivated features that are mainly derived from the linguistic levels syntax, lexicon and morphology.

We trained several linear regression models to find the correlation between rating and these linguistic properties of text. This resulted in a set of 20 features, selected from different linguistic levels, with an error value of 0.2.
The dataset is openly available at the DFKI Github Repository[4] and can be used for further work on subjective text complexity studies or similar research questions.

6. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the projects AuTexx (01IS17043) and vALID (01GP1903A). Moreover, we would like to thank DATEV eG and Prof. Dr. Andreas Both (Head of Research, DATEV) for providing data and helping to conduct the expert-experiments.

7. Bibliographical References

Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., and Ebling, S. (2020). A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France, May. European Language Resources Association.

Broda, B., Ogrodniczuk, M., and Niton, B. (2014). Measuring Readability of Polish Texts: Baseline Experiments. In *Proceedings of the 9th Conference on Language Resources and Evaluation*, pages 573–580, Reykjavik, Iceland, May.

Chall, J. and Dale, E. (1995). *Readability Revisited, the New Dale-Chall Readability Formula*. MA: Bookline Books, Cambridge.

Chatzipanagiotidis, S., Giagkou, M., and Meurers, D. (2021). Broad Linguistic Complexity Analysis for Greek Readability Classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–58. ACL, April.

Dale, E. and Chall, J. S. (1949). The Concept of Readability. *Elementary English*, 26(1):19–26.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., and Macken, L. (2014). Using the Crowd for Readability Prediction. *Natural Language Engineering*, 20(3):293–325, July.

DuBay, W. H. (2004). *The Principles of Readability.* Impact Information, Costa Mesa, California.

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

DuBay, W. (2007). *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, Costa Mesa, California.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 276–284, Beijing, China.

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

Gala, N., Tack, A., Javourey-Drevet, L., François, T., and Ziegler, J. C. (2020). Aector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France, May. European Language Resources Association.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability Classification for German using Lexical, Syntactic, and Morphological Features. In *Proceedings of COLING 2012*, pages 1063–1080, India, Mumbai, December.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., and Mille, S. (2020). Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of The 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland, December. Association for Computational Linguistics.

Imperial, J. M. and Ong, E. (2020). Exploring Hybrid Linguistic Feature Sets to Measure Filipino Text Readability. In *2020 International Conference on Asian Language Processing (IALP)*, pages 175–180. IEEE.

Iskender, N., Polzehl, T., and Möller, S. (2020). Crowdsourcing Versus the Laboratory: Towards Crowd-Based Linguistic Text Quality Assessment of Query-Based Extractive Summarization. In *Proceedings of the Conference on Digital Curation Technologies*, pages 1–16, February.

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., and Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 546–554, Beijing, China, August.

Kincaid, J. P., Fishburne, R. P. J., Rogers, R. L., and Chrissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Millington, TN.

Loughran, T. and McDonald, B. (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance*, 69(4):1643–1671.

Lu, X. (2010). Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Mclaughlin, G. H. (1969). SMOG Grading - a New Readability Formula. *The Journal of Reading*, (22):639–646.

Möller, S. and Raake, A. (2014). *Quality of experience: advanced concepts, applications and methods*. Springer.

Mulyanti, W. and Soeharto, P. (2020). Text Complexity in English Textbooks for Junior High School: A Systemic Functional Perspective. In *Proceedings of the Twelfth Conference on Applied Linguistics (CONAPLIN 2019)*, pages 217–222, Jan.

[4] https://github.com/DFKI-NLP/subjective_text_complexity_corpus

Naderi, B., Mohtaj, S., Ensikat, K., and Möller, S. (2019). Subjective Assessment of Text Complexity: A Dataset for German Language. *arXiv:1904.07733 [cs]*, April.

Naderi, B. (2018). *Motivation of Workers on Microtask crowdsourcing Platforms*. T-Labs Series in Telecommunication Services. Springer.

Nassiri, N., Lakhouaja, A., and Cavalli-Sforza, V. (2018). Arabic Readability Assessment for Foreign Language Learners. In *International Conference on Applications of Natural Language to Information Systems*, pages 480–488. Springer.

Petersen, S. E. and Ostendorf, M. (2009). A Machine Learning Approach to Reading Level Assessment. *Computer Speech & Language*, 23(1):89–106, January.

Pitler, E. and Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 186, Honolulu, Hawaii. Association for Computational Linguistics.

Schiller, A., Teufel, S., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Universität Tübingen, Seminar für Sprachwissenschaft.

Schwarm, S. E. and Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

Severance, S. J. and Cohen, K. B. (2015). Measuring the Readability of Medical Research Journal Abstracts. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing*, pages 127–133, Beijing, China, July.

Vajjala, S. and Lucic, I. (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, Florence, Italy. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 163–173, Montréal, Canada, June.

vor der Brück, T. and Hartrumpf, S. (2007). A Semantically Oriented Readability Checker for German. In *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznan, Poland, October.

Zamanian, M. and Heydari, P. (2012). Readability of Texts: State of the Art. *Theory and Practice in Language Studies*, 2(1):43–53, January.

8. Language Resource References

DeReWo. (2013). *Korpusbasierte Wortgrundformenliste DeReWo, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation*.