A Dataset for Subjective Assessment of German Text Complexity

Salar Mohtaj[1], **Babak Naderi**[1], **Kaspar Ensikat**[1], **and Sebastian Möller**[1,2]
[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]DFKI Projektbüro Berlin, Berlin, Germany
{salar.mohtaj|babak.naderi|sebastian.moeller} @ tu-berlin.de
{ensikat} @ campus.tu-berlin.de

Abstract

This paper presents TextComplexityDE, a text readability assessment dataset consisting of 1000 sentences in German that taken from 23 Wikipedia articles. The corpus could be used for developing text complexity predictors and automatic German text simplification. Text complexity predictor models have diverse applications such as choosing appropriate reading materials for people with intellectual disabilities. The dataset includes subjective assessment of different text-complexity aspects provided by German learners in level A to C. In addition, it contains manual simplification of 250 of those sentences provided by native speakers and subjective assessment of the simplified sentences by participants from the target group.

1 Introduction

Text is a major medium for transforming information in daily human communication. Individuals with different backgrounds face various challenges when comprehending texts written in a complex style (Saggion, 2017). Moreover, the complexity of text can influence its readability and understandability as well as reader's decision making. It has been shown in the domain of micro-task crowd working that the complexity of a task's description and instruction influences the workers' expected workload and consequently affect their decision on whether performing the micro-task or not (Naderi, 2018).

Text complexity is defined as a metric that determines how challenging is a text for a reader (Initiative et al., 2010). It is also described as the sum of all text elements that affect the readers understanding, reading speed and level of interest in the material[1] (Dale and Chall, 1948). It influences the task load and consequently the Quality of Experience (QoE). Readability assessment has diverse use cases and applications, such as helping to choose appropriate learning material for second language learners and people with disabilities (Aluisio et al., 2010). Moreover, It could be used to provide immediate feedback to authors and encourage them to improve the comprehensibility of their text.

Research in text complexity assessment and simplification dates back to the late 1940s. Since then, in the last two decades interest in such systems has grown especially and linguists developed guidelines for clear writing (De Clercq et al., 2014). Researchers attempt to identify text complexity to determine whether 1) a text needs simplification and 2) the text is suitable for a target group (Hancke et al., 2012).

Manual simplification is often performed for text consumed by second-language learners, mostly by modifying the vocabulary (reducing lexical complexity by replacing sophisticated words), syntax (reduced number of constituents per sentence) and improving cohesion (lexical and semantic co-reference) (Crossley and McNamara, 2008; Simensen, 1987; Young, 1999). In recent years automatic text simplification became an important research area in Natural Language Processing (NLP).

Different factors may effect the complexity of text for readers. From the lexical perspective, use of infrequent and non-familiar words, technical terminology and abstract concepts tend to increase the difficulty of the text (Temnikova, 2012). Readers tend to struggle with issues at the syntac-

[1] This definition is very close to the definition of text complexity. As the later is a highly disputed term in linguistics we consider both to be synonym in this paper. For detailed discussion on text complexity see (Vulanović, 2007)

tic level, such as long sentences and convoluted syntax which tend to cause processing difficulties (Harley, 2013).

Most research in text simplification has been done on English and Spanish. For German language, there are two guidelines for simplifying text for two different target groups. The *Einfache Sprache* (easy language) is a convention for targeting readers with weaknesses in reading and writing or those learning German as a second language. It tries to improve readability of text and make it accessible for a broader audience (Kellermann, 2014). The *Leichte Sprache* (plain language) is another convention specifically designed for those with learning and comprehension disabilities. It establishes very strict rules including short main clauses, usage of very common vocabulary and in general avoidance of more complex features of a written text.

The readability score is mostly measured using quantitative features. By the 1980s, there were about 200 formulas and over a thousand studies on the readability formulas verifying their strong theoretical and statistical validity (DuBay, 2004). The Flesch Reading Ease (FRE) score[1], Flesch-Kincaid readability and Gunning Fog Index are the most important and prevalent formulas. They all use measures of average sentence length and average syllables per word for calculations. However, existing formulas vary to a strong degree in their scores even when applied to the same material (Mailloux et al., 1995).

In this paper we proposed TextComplexityDE[2], a dataset for developing and evaluating text complexity predictor models. The proposed dataset could also be used to develop automatic German text simplification tools. It consisting of 1000 sentences in German that taken from 23 Wikipedia articles in 3 different article-genres.

The rest of this paper is organized as follows. Next section contains some description about the available datasets for text readability. In section 3, we explain the dataset structure and how we collected and evaluated the ratings. Obtained ratings are briefly explored in this section as well. In Section 4, we describe the process of manual simplification and finally in section 5, we discuss our findings and present implications for future works.

2 Existing Datasets

Several text readability assessment corpora are already collected for different languages. However, most corpora focus only on article level, i.e. either one readability score assigned to the entire article (e.g. (Schwarm and Ostendorf, 2005) for English containing 2500 articles), or articles are classified to normal or simple (e.g. PWKP data set containing Wikipedia articles and their corresponding simple Wikipedia article (Zhu et al., 2010)).

For the German language, Klaper et al. collected a parallel corpus of German text for normal and plain German by extracting articles from five websites which offer articles in both original and plain language (Klaper et al., 2013). Similarly, Hancke et al. collected articles from two websites, one with the original text and the other, articles with same topic but written for teenager audience (Hancke et al., 2012). In both cases, text complexity is only differentiated between two levels. Other researches used indirect measurements techniques like using eye-tracking, context questions, or measurements of effort to estimate the text complexity (Jekat et al., 2018).

To the best of our knowledge, there is no corpus available for German language containing subjective assessments of text complexity, or parallel simplification of text in sentence level. Therefore, a newly annotated corpus is required for future development of automatic text complexity assessment and automatic simplification. Such a corpus should contain subjective ratings (at least) in sentence level and parallel simplification of the original text in sentence level. The main contributions of this paper are as follows:

- Presenting a German text readability corpus containing 1000 sentences with their subjective complexity assessment from language learner group
- Developing a scale for collecting subjective complexity ratings
- Presenting a parallel corpus with the original and simplified version of 250 sentences

3 Subjective Assessment of Text Complexity

To compile our German text readability corpus, we collect subjective ratings in the sentence level and focus on German learners as our target group. To

[1] Amstad adjusted it for German (Amstad, 1978)

[2] Temporal URL: http://tiny.cc/mq643y

collect reliable and valid data, (Recommendation, 2018a,b) standards are adopted in the study design and data screening process to perform Absolute Category Ratings (ACR). The detailed description of the corpus generation process is presented in this section.

3.1 Source Text

Sentences contained in the dataset were collected from two sources; the German Wikipedia[1] (1000 sentences) and 100 sentences from the Leichte Sprache (Simple language) dataset (Klaper et al., 2013).

The reason to choose Wikipedia as the main resource is that its articles are written for general native audiences and not yet specifically tailored for those who are still in the process of learning. Moreover, Wikipedia articles are mostly written by several volunteers therefore we expect they cover a wide range of linguistic levels and writing styles.

Sentences from the Leichte Sprache were used as Gold Standard Questions (Naderi, 2018) as indicator for the quality of data collected in a rating session. We took 23 articles from three domains (history, society and science) from Wikipedia and two articles from the Leichte Sprache.

3.2 Development of Scale

We conducted a pilot study to determine relevant dimensions of text complexity that can be captured within the subjective assessment. An initial item pool with 11 questions was developed and reviewed by a linguistic expert (cf. Table 1).

Within a pilot study, 100 sentences were assessed by crowd workers. We created 20 crowdsourcing jobs in the ClickWorker[2] micro-task crowdsourcing platform. In each job, participants assessed five sentences (4 from Wikipedia pool and one from Leichte Sprache) by answering the 11 questions from the item pool. In total, ten different workers rated each sentence. Overall, 77 German learners participated in the study (submission from native speakers were discarded).

3.2.1 Data Screening

Different strategies have been used to refine the submitted data by the crowd workers.

Firstly, submissions from workers with unreasonable completion times or unrealistic answer to the gold standard question were removed. Moreover, responses were evaluated against unexpected patterns in ratings (i.e. no variance or potential outliers). Uni-variate outliers were identified in item level by calculating the standardized scores (absolute z-score larger than 3.29 considered to be a potential outlier (Naderi, 2018)). Submissions with more than one potential outliers were removed from the final dataset. Finally, 122 answer packages (i.e. 610 ratings) were accepted.

3.2.2 Evaluation

The Mean Opinion Score (MOS) value for each item was calculated per sentence using the accepted answer packages. Ten sentences were removed from the final data, since there were less than five votes available for them. Items were investigated by calculating the Cronbach's α value and the internal consistency, assuming that all items express the same construct i.e. text complexity. The α value of .996 is achieved by removing 4 items (i.e. item 3, 5, 8 and 10). Next, Principal Component Analysis (PCA) was performed which leads to extract two factors; complexity (item 1) was the dominant item loading on the first factor and understandability (item 7) was loading on the second factor.

3.3 Data Collection Procedure

An online survey system was created to collect the subjective assessment of 1000 sentences using three items each rated on a 7-point Likert Scale. A survey session consist of training and rating sections. Each session was started by three demographic questions followed by a training section and finally the assessment of ten sentences. Participants provide their age, education and German language level according to the CEFR.

The training section was containing three sentences which participants needed to rate on the same scale as the main section. The sentences in the training section were constant and represent very easy, average and very complex sentences. Afterward, participants rated complexity, understandability and lexical difficulty[3] of ten sentences by answering to the following questions on 7-point Likert scales:

- **Complexity:** *How do you rate the complexity of the sentence?* Scale from *very easy (1)* to

[1] http://de.wikipedia.org/

[2] https://www.clickworker.com/

[3] This item was included as we aim to investigate it in future.

Table 1: Initial item set used in the pilot study

	Item
1	How do you rate the overall complexity of the sentence?
2	How difficult was it for you to read this sentence?
3	How familiar are you with the topic of the article?
4	How difficult would it be to translate this sentence into your native language?
5	How many different ways can this sentence be interpreted?
6	How difficult would it be to explain this sentence to another person?
7	How well did you understand the sentence?
8	How many words in this sentence are unfamiliar to you?
9	Take a look at the hardest words contained in the sentence. How difficult is it for you to understand those words?
10	How many words in this sentence have multiple interpretations?
11	How do you rate the complexity of the syntactical structure of the sentence?

very complex (7).

- **Understandability:** *How well were you able to understand the sentence?* Scale from *fully understood (1)* to *didnt understand at all (7).*
- **Lexical difficulty:** *Regarding the hardest words in the sentence: How difficult is it to you, to understand these words?* Scale from *very easy (1)* to *very difficult (7).*

Users could participate in the survey as many times as they wanted and the system was designed to avoid a same sentence to be assigned to the same participant on their return.

3.3.1 Participants

We aimed to collect at least ten votes per sentence. From 369 participants in the study, 267 reported a German language level between A and B. In total, the survey was completed 1322 times. Out of those 1065 were provided valid ratings from German language learners resulting in 10650 valid sentence ratings split across the 1000 sentences. Participants were recruited from three channels:

- Paid crowdsourcing (16% of valid answers)
- Volunteers[1] (21% of valid answers)
- Laboratory study[2] (63% of valid ratings)

The third group were recruited by contacting students in German language courses (B level) offered by the Technical University of Berlin. They were asked to rate sentences in intervals of 20 Minutes with short breaks in between. Participants were compensated for their efforts and on average completed the survey 20 times (rating 200 sentences). The demographic information of participants is presneted in Table 2.

Table 2: The demographic information of participants

Age	<25	23%
	25 - 35	42%
	>35	35%
German skill	A	13%
	B	60%
	C	27%
Native language	Spanish	12%
	Russian	8%
	Arabic	7%
	English	6%
	Others	67%

3.3.2 Absolute Category Ratings

Following the data screening process explained in Section 3.2.1, 5 to 18 valid ratings for each sentence remained in the dataset. For each sentence MOS, standard deviation and 95% Confidence Intervals of each dimension are reported in the dataset. Fig. 1 illustrates the distribution of Mean opinion score (MOS) values. As expected sentences from Wikipedia are more complex (M_{MOS} = 3.22) than the sentences from the

[1] Through 87 Facebook groups organized by German learners.

[2] 33 German learners laboratory sessions from 1 to 1:30 hours

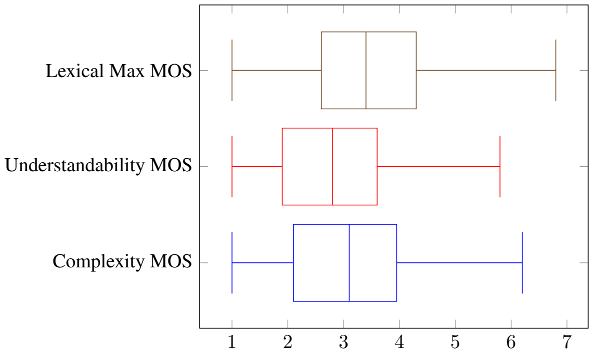

Figure 1: Distribution of MOS of Complexity, Understandability and Lexical difficulty ($N = 1000$)

plain language dataset ($M_{MOS} = 1.2$). In addition, there are strong significant correlations between the three dimensions: complexity has a correlation of .896 with understandability and .905 with lexical difficulty. Also, Understandability has a correlation of .935 with lexical difficulty.

In addition, Amstad's adaptation of FRE score (Amstad, 1978) has been used to calculate the readability score for both article (n=25) on sentence level. We used the collected ratings to calculate complexity score in article level[1]. The FRE-scores for the two articles written in plain language were 62 and 66 (out of 100). While it could be interpreted as moderately difficult, participants in our study considered them as very easy. Moreover, the FRE-score and average MOS rating of complexity in article level were strongly correlate ($r = .89$, $p < .001$), with a huge intercept when they are normalized (2.6 in 7 point range). Also, both values are moderately correlate ($r = .55$, $p < .001$) in sentence level. Strong disagreement in highest range of FRE-score in our study confirm previous research that the FRE-formula does not perform well at sentence-level (McClure, 1987). Finally, the sentences with highest complexity rating were examined by native speakers. It revealed that, they are either thematically complex even for the native speakers or are written in a convoluted manner.

4 Manual Simplification

Based on the complexity ratings provided in previous chapter, subset of the dataset were selected for manual simplification. 265 sentences with complexity rating above 4 point of MOS and understandability rating above 3.5 MOS were selected for manual simplification. Overall, 659 simplifications of original sentences were collected from 75 native speakers. For 250 out of 265 sentences at least one simplification were provided. In 90 cases native speakers reported that they were unable to simplify the provided sentence.

5 Discussion and Future Work

This work presents a corpus of 1000 sentences in German, includes the complexity, understandability and lexical difficulty. The measures are assessed by a group of language learners participated in subjective studies conducted following best practices in the quality of experience community. Moreover, the corpus contains manual simplifications for 250 sentences, written by native German speakers. It should be noted that subjective ratings refer to the degree that participants perceived a concept. For some aspect of text it might be important to not only measure the perceived degree but also the actual value of concept. For instance, the understandability measurement in this study refers to participants' thought of what they understood from a given text. It may differ from the actual level of understanding for which different assessment methods like content questions should be used. Therefore, researchers should carefully decide which kind of measurement technique to employ depending to the goals of their study. For future work, we would like to compare subjective assessment of text understandability and complexity as explained in this paper with actual understandability (e.g. measured by content questions) and readability (e.g. measured by eye-tracking) scores.

[1] We used average as a very basic model.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

T Amstad. 1978. *Wie verständlich sind unsere Zeitungen?[How readable are our newspapers?].* Ph.D. thesis, Doctoral thesis, Universität Zürich, Switzerland.

Scott A Crossley and Danielle S McNamara. 2008. Assessing l2 reading texts at the intermediate level: An approximate replication of crossley, louwerse, mccarthy & mcnamara (2007). *Language Teaching*, 41(3):409–429.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

William H DuBay. 2004. The principles of readability. *Online Submission*.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. *Proceedings of COLING 2012*, pages 1063–1080.

Trevor A Harley. 2013. *The psychology of language: From data to theory*. Psychology press.

Common Core State Standards Initiative et al. 2010. Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects, appendix a. *Retrieved June*, 1:2010.

Susanne Johanna Jekat, Klaus Schubert, and Martin Kappus. 2018. Barrieren abbauen, sprache gestalten. *Working Papers in Applied Linguistics*.

Gudrun Kellermann. 2014. Leichte und einfache sprache–versuch einer definition. *Aus Politik und Zeitgeschichte*, 64(9-11):7–10.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a german/simple german parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19.

Stephen L Mailloux, Mark E Johnson, Dennis G Fisher, and Timothy J Pettibone. 1995. How reliable is computerized assessment of readability? *Computers in nursing*, 13:221–221.

Glenda M McClure. 1987. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, PC-30(1):12–15.

Babak Naderi. 2018. *Motivation of Workers on Microtask Crowdsourcing Platforms*, 1st edition. Springer Publishing Company, Incorporated.

ITU-T Recommendation. 2018a. Recommendation p.800 : Methods for subjective determination of transmission quality. *International telecommunication union*.

ITU-T Recommendation. 2018b. Recommendation p.808: Subjective evaluation of speech quality with a crowdsourcing approach. *International telecommunication union*.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

Aud Marit Simensen. 1987. Adapted readers: How are they adapted. *Reading in a foreign language*, 4(1):41–57.

Irina Temnikova. 2012. *Text Complexity and Text Simplification in the Crisis Management domain*. Ph.D. thesis, University of Wolverhampton, Wolverhampton, UK.

Relja Vulanović. 2007. On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics*, 20:399–427.

Dolly N Young. 1999. Linguistic simplification of SL reading material: Effective instructional practice? *The Modern Language Journal*, 83(3):350–366.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.