Evaluating Readability Metrics for German Medical Text Simplification

Karen Scholz[1,2] **Markus Wenzel**[1]

[1]Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
[2]University of Bremen, Faculty 3: Mathematics/Computer Science, Germany

{karen.scholz, markus.wenzel}@mevis.fraunhofer.de

Abstract

Clinical reports and scientific health information sources are usually written for medical experts preventing patients from understanding the main messages of these texts. Making them comprehensible for patients is important to enable patients to make informed health decisions. Metrics are required to assess readability and to evaluate text simplification methods. However, research has mainly focused on English medical texts. We collected a set of 18 statistical, part-of-speech-based, syntactic, semantic and fluency metrics from related studies and evaluate their suitability to measure readability of German medical texts. We perform multiple t-tests on technical abstracts from English and German scientific articles and related simplified summaries, respectively. While semantic and fluency metrics can be successfully transferred to German medical texts, multiple statistical, part-of-speech-based, and syntactic metrics behave differently when they are applied to German medical texts requiring careful interpretation.

1 Introduction

Healthcare and medicine has evolved towards a more patient-centered and personalized patient care where patients are encouraged to engage in health decision making. Patient engagement promises to enable more personalized treatment planning and to improve therapy outcomes while reducing the risk of medical errors (Alarifi et al., 2020). Patients need a good understanding of their own health condition and therapeutic options to make informed health decisions. However, due to a lack of personnel, physicians often have only limited time for patient education. In several countries, patients can pre-inform about their examination results before their next appointment by accessing their clinical reports via online portals (Baun et al., 2020; BMG, 2024; Cho et al., 2020; Dercksen and de Vries, 2020), which improves communication between patients and physicians (Woods et al., 2013). The ability of patients to understand provided health information materials and leverage gained knowledge for health decision making is referred to as health literacy (Sørensen et al., 2012). However, while clinical reports and further scientific information sources target medical experts, many patients have problems understanding these texts (Cho et al., 2020; Keselman and Smith, 2012; Rogers et al., 2023). To make clinical reports and medical scientific texts comprehensible for patients, several research studies have investigated automatic methods for medical text simplification (MTS) using classical natural language processing (NLP) (Kloehn et al., 2018; Qenam et al., 2017), and deep learning (DL) approaches (Devaraj et al., 2021; Jeblick et al., 2024; Lyu et al., 2023; Phatak et al., 2022).

To develop suitable MTS methods, readability metrics play a crucial role, which is twofold. First, by measuring language characteristics of medical texts, readability metrics enable to identify language patterns causing comprehension problems. As such, they are required to define the requirements for MTS methods. Second, to evaluate these methods, many studies rely on human evaluation studies where medical experts and laypeople rate the readability of simplified texts (Jeblick et al., 2024; Lyu et al., 2023; Moramarco et al., 2021). Since human evaluation studies are expensive and their reproducibility is limited, suitable automatic readability metrics are required to increase scalability of MTS evaluation and to ensure comparability of different MTS

Proceedings of the 31st International Conference on Computational Linguistics, pages 6049–6062
January 19–24, 2025.

methods. However, previous studies on readability metrics for MTS mainly focused on English scientific articles (Kauchak et al., 2017; Leroy et al., 2010; Leroy et al., 2013; Mukherjee et al., 2019). To the best of our knowledge, there are only few studies considering non-English medical texts (Mukherjee et al., 2017) and clinical reports (Zeng-Treitler et al., 2007).

Our contributions are as follows: We collect a set of 18 statistical, grammar, semantic, and fluency readability metrics from related studies on readability metrics and MTS. To evaluate whether available metrics are suitable for measuring readability of German medical texts, we validate these metrics on sets of paired technical abstracts and corresponding simplified summaries from English and German scientific articles using multiple t-tests and compare results.

2 Related Work

Several studies have investigated patient comprehension of medical texts and identified that medical expert language prevents laypeople from understanding these texts. Alarifi et al. (2021) collected 659 questions about clinical online portals from online discussion forums. The authors identified that understanding their clinical reports is the main concern of patients as indicated by more than one third of all questions. Rogers et al. (2023) conducted a systematic literature review including studies about patients reporting about their understanding of clinical reports and conclude that problems with respect to language and medical terminology prevent patients from understanding the main messages. Other studies confirm medical terminology, abbreviations, and unclear structure as main difficulties in clinical report language (Keselman et al., 2007). Gunn et al. (2017) conducted a user study with laypeople rating comprehensibility of imaging reports. The authors determined a median comprehensibility of 2.5 on a 5-point Likert scale independent of demographic features. Only participants with prior experience in reading clinical reports had significantly better understanding. Cho et al. (2020) compared comprehensibility of clinical reports written in expert language and corresponding simplified summaries and found that providing simplified versions improves comprehensibility. Zowalla et al. (2023) analyzed readability of German health-related web pages showing that readability is low.

To make clinical reports and other health information sources accessible to patients without increasing the workload on medical staff, researchers have developed automatic MTS methods. While former MTS methods mainly leveraged classical NLP methods in combination with terminologies and lexica (Kloehn et al., 2018; Qenam et al., 2017), recent approaches increasingly use DL models (Devaraj et al., 2021; Phatak et al., 2022) including large language models (LLMs) (Jeblick et al., 2024; Lyu et al., 2023). To assess the readability of medical texts, to identify difficult language patterns, and to evaluate the performance of MTS methods, suitable readability metrics are required. Many studies rely on manual evaluation strategies (Jeblick et al., 2024; Lyu et al., 2023; Moramarco et al., 2021). However, results obtained this way are hardly reproducible, which compromises comparison of MTS methods. Thus, the availability of suitable automatic readability metrics is important. Previous studies mainly focused on English scientific texts. Leroy et al. (2013) validated term frequency, i.e. the frequency of occurrence of a term in common language, as a metric for semantic readability of English scientific medical texts. Replacing low-frequency with higher-frequency terms showed to significantly improve readability in a user study. Mukherjee et al. (2019) found that lexical chain-based metrics are sufficient for measuring fluency in English medical scientific texts showing that difficult texts suffer from more topic changes and intersecting threads. Another study by Zeng-Treitler et al. (2007) investigated the suitability of lexical, syntactic, and semantic readability metrics to characterize language in clinical reports. The earlier work by Mukherjee et al. (2017) is one of the few studies considering non-English texts. The authors validated statistical, grammar, and semantic metrics on Spanish medical scientific texts. Naderi et al. (2019) used 20 statistical, lexical, and morphological metrics to predict readability of German texts finding that morphological metrics perform better than statistical and lexical metrics. Similarly, Weiss and Meurers (2022) used a set of 373 lexical, syntactic, morphological, and cohesion metrics among others to predict readability in German texts showing that these outperform traditional readability formulas. However, both studies do not consider medical texts.

3 Methods and Materials

To evaluate the suitability of available readability metrics for German medical texts, we collect a set of 18 readability metrics from related studies on readability metrics and MTS. The metrics are introduced in Section 3.2. We validate collected readability metrics on pairs of technical abstracts and related simplified summaries from English and German scientific articles, respectively. Similar to Mukherjee et al. (2017), we perform multiple significance tests to identify those readability metrics suitable to differentiate between technical abstracts and simplified summaries and compare results from both languages. The datasets are described in Section 3.1. Analysis methods are presented in Section 3.3.

3.1 Datasets

Scientific articles are obtained from the Cochrane Database of Systematic Reviews[1] where review articles summarizing current scientific studies on medical topics are published. All articles contain a technical abstract and a plain language summary (PLS). Technical abstracts summarize the main findings of the full article using expert language. To make scientific information accessible to laypeople, a PLS is provided. PLSs have to follow a respective writing guide provided by the Cochrane Library Editorial Board and summarize the main findings of the full article using plain language (Pitcher et al., 2022). Devaraj et al. (2021) and Joseph et al. (2023) already created datasets to train MTS models from paired technical abstracts and PLSs from the Cochrane Database showing that there is a significant overlap in terms of content. Articles are available in English by default. They are translated by volunteering native speakers, professional translators, and Cochrane-internal translator teams using machine translation in combination with human post-editing into other languages including German (Deppe, 2024). We filtered the Cochrane Database by Cochrane Topics "*Cancer*" and "*Heart & circulation*" and extracted all English and German technical articles and PLSs from reviews for which German translations are available for both, technical abstracts and PLSs. As a result, we obtain an English and a German dataset both consisting of 50 pairs of technical abstracts and related PLSs from 50 articles in total.

3.2 Readability Metrics

We implemented 18 readability metrics in total. We categorize them into statistical, grammar, which comprise syntactic and part-of-speech- (POS)-based metrics, semantic, and fluency metrics.

Formally, we consider dataset $D = \{d_l\}_{l=1}^{N}$ with N text documents. All text documents $d \in D$ are preprocessed using *spaCy* (Honnibal et al., 2020), an open-source NLP programming library for Python, before metrics are calculated for each text. Preprocessing includes tokenization, POS-tagging, dependency parsing, sentence and syllables splitting. We denote the sequence of sentences in text document d of length L_{s_d} by $s_d = (s_{id})_{i=1}^{L_{s_d}}$, which is obtained from sentence splitting. Similarly, we denote the sequence of words in text document d of length L_{w_d} by $w_d = (w_{jd})_{j=1}^{L_{w_d}}$ and the sequence of words in sentence s_{id} of length $L_{w_{s_{id}}}$ by $w_{s_{id}} = (w_{js_{id}})_{j=1}^{L_{w_{s_{id}}}}$. These are obtained from tokenization. Let $p_{s_{id}} = (p_{js_{id}})_{j=1}^{L_{w_{s_{id}}}}$ denote the sequence of POS categories obtained by assigning each word in s_{id} its POS category. Similar to Kauchak et al. (2017), we obtain the dependency parse trees for each sentence from the dependency parser. We denote the dependency parse tree of sentence s_{id} by $T_{s_{id}} = (V_{s_{id}}, E_{s_{id}})$, which is a directed rooted tree. We label the tree nodes $V_{s_{id}}$ with the POS categories of the words in sentence s_{id}, so $V_{s_{id}} = \{p_{js_{id}}\}_{j=0}^{L_{w_{s_{id}}}}$ with $p_{0s_{id}}$ being the root node. There is a directed edge $(p_{js_{id}}, p_{ks_{id}}) \in E_{s_{id}}$ if word $w_{ks_{id}}$ syntactically depends on word $w_{js_{id}}$ as indicated by the dependency parser. Let $\text{depth}(p_{js_{id}}, T_{s_{id}})$ denote the number of steps from $p_{0s_{id}}$ to $p_{js_{id}}$. Based on that, the n^{th}-level dependency parse tree $T_{\leq n, s_{id}}$ denotes the parse tree containing only nodes $p_{js_{id}}$ with $\text{depth}(p_{js_{id}}, T_{s_{id}}) \leq n$ (Kauchak et al. 2017). We finally denote the sequence of syllables in text document d of length L_{b_d} by $b_d = (b_{ld})_{l=1}^{L_{b_d}}$ resulting from syllables splitting.

Statistical metrics provide surface-level statistics over a given text such as paragraph, sentence, and word lengths. Here, we included three statistical metrics: **(1) Average sentence**

[1] https://cochranelibrary.com

length (ASL) is measured by the average number of words per sentence (Equation 1):

$$\mathrm{ASL}_d = \frac{L_{w_d}}{L_{s_d}} \quad (1)$$

This is motivated by Mukherjee et al. (2017), who have already proven that average sentence length is a suitable metric to measure lexical complexity in Spanish texts. **(2) Average word length** (AWL) is estimated by the average number of syllables per word (Equation 2):

$$\mathrm{AWL}_d = \frac{L_{b_d}}{L_{w_d}} \quad (2)$$

We further consider **(3) Flesch-Kincaid Grade Level** (FKGL) (Kincaid et al., 1975), which is a readability formula estimating the US-American grade level required to understand a text. FKGL considers both, average word and sentence lengths, as shown in Equation 3:

$$\mathrm{FKGL}_d = 0.39\ \mathrm{ASL}_d + 11.8\ \mathrm{AWL}_d - 15.59 \quad (3)$$

FKGL has been applied by several MTS studies to evaluate readability (Basu et al., 2023; Phatak et al., 2022; Yang et al., 2023).

POS-based metrics measure grammatical complexity of texts by calculating the frequency distribution of POS categories, i.e. the proportion of words assigned to a certain set of POS categories to the total number of words in the text (Mukherjee et al. 2017). Let TS denote this set of target POS categories and $w_{TS,d} = (w_{jd} | p_{jd} \in TS)_{j=1}^{L_{w_d}}$ the list of words which is assigned a POS from TS of length $L_{w_{TS,d}}$. Then, the POS proportion (POSPRP) is calculated according to Equation 4:

$$\mathrm{POSPRP}_{TS,d} = \frac{L_{w_{TS,d}}}{L_{w_d}} \quad (4)$$

Mukherjee et al. (2017) identified that difficult Spanish medical texts exhibit significantly more nouns, negations, and adjectives, while simple texts use more numbers. Leroy et al. (2010) found that function words are indicative for readability of English texts. Based on that, we decided to include **(4) noun ratio**, **(5) adjective ratio**, **(6) function word ratio**, **(7) negation ratio** and **(8) numbers ratio** in our study. Hereby, we consider determiners, auxiliaries, adpositions, modals, and pronouns as function words. We calculate these metrics based on Equation 4 with metric-specific sets of target POS categories. We provide an overview of the applied target POS categories for each metric, respectively, in Appendix B.

Syntactic metrics aim to measure complexity of language by assessing complexity of sentence structures and word compositions. We implemented four syntactic metrics in total. **(9) Average edit distance** and **(10) grammar frequency** measure variability of sentence structures as proposed by Mukherjee et al. (2017). For English medical texts, Kauchak et al. (2017) have already shown that sentence structure has an effect on text understanding. Similar to Mukherjee et al. (2017), we obtain the average edit distance by calculating the edit distances of the sentence structures of all adjacent sentence pairs of a text and average them. Therefore, we represent the sentence structure of sentence s_{id} by its associated sequence of POS categories $p_{s_{id}}$. The average edit distance (AED) is obtained according to Equation 5, where $\mathrm{dist}(p_{s_{id}}, p_{s_{i+1d}})$ denotes the Levenshtein distance.

$$\mathrm{AED}_d = \frac{1}{L_{s_d}-1} \sum_{i=1}^{L_{s_d}-1} \frac{\mathrm{dist}(p_{s_{id}}, p_{s_{i+1d}})}{\max(L_{w_{s_{id}}}, L_{w_{s_{i+1d}}})} \quad (5)$$

To ensure comparability, we normalize Levenshtein distances with respect to the sentence lengths. Grammar frequency is obtained from the proportion of different sentence structures to the total number of sentences in a text (Mukherjee et al. 2017). Different from average edit distance, we obtain the sentence structures from the third-level dependency parse tree of each sentence as proposed by Kauchak et al. (2017). Let $T_{\leq 3,d} = \{T_{\leq 3,s_{id}}\}_{i=1}^{L_{s_d}}$ denote the set of unique parse trees obtained for text document d. Grammar frequency (GF) is calculated according to Equation 6:

$$\mathrm{GF}_d = \frac{|T_{\leq 3,d}|}{L_{s_d}} \quad (6)$$

We further consider **(11) maximum dependency tree depth** (MDTD) as proposed by Menta and Garcia-Serrano (2022) to measure sentence structure complexity. We calculate the maximum depths of the dependency parse trees for all sentences of a text and average them (Equation 7):

$$\mathrm{MDTD}_d = \frac{1}{L_{s_d}} \sum_{i=1}^{L_{s_d}} \max((\mathrm{depth}(p_{js_{id}}, T_{s_{id}}))_{j=1}^{L_{w_{s_{id}}}}) \quad (7)$$

We further consider **(12) noun phrase complexity** (Leroy et al. 2010) which denotes the number of words of a multi-word phrase with a noun as its base. Leroy et al. (2010) showed that using

complex noun phrases is characteristic of difficult English medical texts. Let $n_d = (n_{ld})_{l=1}^{L_{n_d}}$ denote the list of noun phrases in text document d. Every noun phrase n_{ld} is a list of words $w_{n_{ld}} = (w_{jn_{ld}})_{j=1}^{L_{w_{n_{ld}}}}$ of length $L_{w_{n_{ld}}}$. The average noun phrase complexity (ANPC) is calculated according to Equation 8:

$$\text{ANPC}_d = \frac{1}{L_{n_d}} \sum_{l=1}^{L_{n_d}} L_{w_{n_{ld}}} \quad (8)$$

Semantic metrics are used to measure how difficult it is for readers to understand the meaning of terms in a text. A prominent semantic readability metric constitutes **(13) average term frequency** (ATF). Based on the assumption that simple words are more frequently used in common language than difficult words, term frequency has been successfully used as an estimate for semantic readability for English medical texts by Leroy et al. (2013). We calculate the term frequency $\text{tf}(w_{jd})$ for each word of a text document using the Python library *wordfreq* (Speer, 2022) and average obtained frequencies (Equation 9):

$$\text{ATF}_d = \frac{1}{L_{w_d}} \sum_{j=1}^{L_{w_d}} \text{tf}(w_{jd}) \quad (9)$$

Another approach for measuring semantic readability is proposed by Menta and Garcia-Serrano (2022) called **(14) Language Model Fill-Mask** (LMFM). LMFM leverages a masked language model (MLM) to predict words, which are randomly masked out of the input text. By ranking all words of the MLM vocabulary according to their prediction probabilities, the semantic readability of the original word is estimated by its position in this ranking. We mask and predict words sentence-wise using a masking ratio of 15%. Let VO_{MLM} denote the set of tokens in the MLM's vocabulary. Let $m_d = (m_{ld})_{l=1}^{L_{m_d}}$ denote the list of masked words in text document d. Let $\text{rank}(m_{ld})$ denote the ranking of word m_{ld} when sorting the output probabilities of the MLM for all tokens in VO_{MLM} in descending order. The obtained ranking positions are normalized with respect to the MLM's vocabulary size. The average LMFM score for text document d is calculated according to Equation 10:

$$\text{LMFM}_d = \frac{1}{L_{m_d}} \sum_{l=1}^{L_{m_d}} \frac{\text{rank}(m_{ld})}{|VO_{\text{MLM}}|} \quad (10)$$

As MLMs, we use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) for English and German BERT[2] for German texts, respectively.

Fluency metrics measure coherence of text. Laban et al. (2021) propose to use **(15) language model (LM) perplexity** to measure fluency. LM perplexity leverages a generative LM optimized to predict the next token given all previous tokens of a tokenized text. We calculate LM perplexity separately for all sentences in a text and average obtained perplexities over the sentences. Let $\mathcal{L}_{\text{LM, CE}}(s_{id})$ denote the cross entropy (CE) loss of an LM on sentence s_{id}, where the expected prediction for each token in sentence s_{id} is its subsequent token. The average LM perplexity (APPL) is calculated according to Equation 11:

$$\text{APPL}_d = \frac{\sum_{i=1}^{L_{s_d}} e^{\mathcal{L}_{\text{LM,CE}}(s_{id})}}{L_{s_d}} \quad (11)$$

As LMs, we use Generative Pre-trained Transformer (GPT)-2 (Radford et al., 2019) and German GPT-2 (Schweter, 2020) for English and German texts, respectively. Mukherjee et al. (2019) propose to use lexical chain-based metrics to measure fluency. The authors denote a lexical chain as a part of a text dealing with a specific topic, where a lexical chain is formed by the sequence of all occurrences of a specific noun in a text. The authors showed that difficult English medical texts have significantly shorter chains, i.e. topics change more frequently, and more intersecting chains. Similar to Mukherjee et al. (2019) we calculate **(16) average chain length** as the average number of nouns forming a lexical chain. Let $c_d = \{c_{ld}\}_{l=1}^{L_{c_d}}$ denote the set of lexical chains in text document d where each chain $c_{ld} \in c_d$ is a sequence of identical nouns $w_{c_{ld}} = (w_{jc_{ld}})_{j=1}^{L_{w_{c_{ld}}}}$ of length $L_{w_{c_{ld}}}$. The average chain length ACL is calculated according to Equation 12:

$$\text{ACL}_d = \frac{1}{L_{w_d}} \left(\frac{\sum_{l=1}^{L_{c_d}} L_{w_{c_{ld}}}}{L_{c_d}} \right) \quad (12)$$

We further calculate **(17) the number of cross chains** as the total number of chains in a text

[2] https://huggingface.co/dbmdz/bert-base-german-uncased

crossing or partly overlapping each other. Similar to Mukherjee et al. (2019), both metrics are normalized with respect to the text length L_{w_d}. Lastly, we use **(18) next sentence prediction (NSP) score** to estimate coherence. NSP score uses BERT-based MLMs, which have been pre-trained on NSP task (Devlin et al., 2019) to predict for all adjacent sentence pairs in a text whether they are adjacent sentences. Let $P_{\text{NSP}}(\text{next}|s_{id}, s_{i+1d})$ denote the predicted probability of an MLM that sentence s_{i+1d} directly follows sentence s_{id} in text document d and $P_{\text{NSP}}(\text{not next}|s_{id}, s_{i+1d})$ the probability that s_{i+1d} does not follow s_{id}. The average NSP score (ANSP) results from the proportion of pairs of adjacent sentences in text d for which adjacency is confirmed by the MLM as shown in Equations 13 and 14:

$$\text{ANSP}_d = \frac{1}{L_{s_d}-1}\sum_{i=1}^{L_{s_d}-1} \text{nsp}(s_{id}, s_{i+1d}) \tag{13}$$

$$\text{nsp}(s_{id}, s_{i+1d}) = \begin{cases} 1, & P(\text{next}|s_{id}, s_{i+1d}) > \\ & P(\text{not next}|s_{id}, s_{i+1d}) \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

We use BERT (Devlin et al., 2019) and German BERT as MLMs for English and German texts, respectively.

3.3 Analysis Methods

We calculate the readability metrics from Section 3.2 for technical abstracts and PLSs from the extracted English and German datasets from Section 3.1. We perform paired sample t-tests between technical abstracts and corresponding PLSs for all readability metrics, to identify which metrics are suitable to measure readability of medical texts. We apply Bonferroni correction with a significance level of 0.05 to accommodate for error accumulation due to multiple testing resulting in a corrected significance level of 0.003 (rounded to three decimals places) for 18 metrics. Paired sample t-tests are performed separately for English and German datasets.

4 Results

4.1 English Medical Texts

T-test results for English medical texts (Table 1) reveal that two third of the readability metrics differ significantly between technical abstracts and PLSs. In terms of **statistical metrics**, technical abstracts have significantly longer sentences with more than four words more on average compared to their simplified counterparts. Similarly, FKGL is significantly higher in technical abstracts compared to PLSs. However, we do not observe a significant difference in word lengths as measured by the number of syllables per word, which implies that the decrease in FKGL results from the reduction of sentence lengths.

POS-based metrics show that technical abstracts use significantly more nouns and numbers than PLSs. In contrast, PLSs use significantly more function words and negations than technical abstracts. Only for the proportion of adjectives, we do not observe significant differences. **Syntactic metrics** including average edit distance and grammar frequency do not indicate significant differences in terms of sentence structure variability between English technical abstracts and PLSs. However, depths of dependency parse trees from sentences in PLSs are significantly lower compared to those from sentences in technical abstracts implying that sentence structure complexity is reduced in PLSs which might also be caused by the reduction of sentence lengths. Noun phrases in technical abstracts are longer than in PLSs, which is also significant.

Semantic metrics including term frequency and LMFM show that PLSs use terms which occur more frequently in common language and which are easier to predict by BERT MLM than technical abstracts. This difference is significant as indicated by both metrics. These findings are in line with previous studies on English medical texts (Leroy et al., 2013) and suggest that PLSs are easier to understand on the semantic level than their technical counterparts.

Fluency metrics show that there is no significant difference between technical abstracts and PLSs in terms of fluency at sentence level as indicated by LM perplexity. Considering fluency at full text level, lexical chain metrics show that technical abstracts have shorter lexical chains and more intersecting chains as measured by average lexical chain length and average number of cross chains, respectively. Both metrics differ significantly between technical abstracts and PLSs. This is in line with the findings of Mukherjee et al. (2019) suggesting that topics in technical abstracts

Metric Type	Metric Name	Technical Abstract Mean	Plain Language Summary (PLS) Mean	p-value
Statistical	Words per sentence	25.713	21.332	**0.000**
	Syllables per word	1.574	1.588	0.409
	FKGL	13.007	11.472	**0.000**
POS-based	Noun ratio	0.398	0.330	**0.000**
	Adjective ratio	0.092	0.010	0.034
	Function word ratio	0.222	0.282	**0.000**
	Numbers ratio	0.091	0.027	**0.000**
	Negations ratio	0.003	0.005	**0.000**
Syntactic	Grammar frequency	0.988	0.991	0.489
	Average edit distance	0.743	0.747	0.503
	Dependency tree depth	7.874	7.467	**0.000**
	Noun phrase complexity	1.958	1.865	**0.000**
Semantic	Term frequency	0.006	0.007	**0.000**
	LMFM	0.281	0.185	**0.000**
Fluency	LM perplexity	79.601	66.731	0.004
	Average lexical chain length	0.006	0.011	**0.000**
	Average number of cross chains	1.043	0.502	**0.000**
	NSP score	0.992	0.972	0.008

Table 1: Means and p-values (p) obtained from paired sample t-tests for English technical abstracts and related plain language summaries (PLSs). Significant differences are highlighted in bold given a significance level of $p < 0.003$ after Bonferroni correction.

are discussed more briefly and thus change more frequently, while at the same time, the higher number of intersecting chains indicates that language complexity additionally arises from the fact that multiple threads are parallelly discussed and brought into context. No significant differences are observed for NSP score.

4.2 German Medical Texts

While we observed significant differences for two third of the metrics (twelve out of 18) in English texts, in German texts significant differences are observed even for 13 out of the 18 metrics according to our t-test results (Table 2). **Statistical metrics** indicate that sentences in German PLSs tend to be longer than sentences in technical abstracts. This is indicated by the average number of words per sentence, which is about one word higher on average in PLSs compared to technical abstracts. However, this is not significant but in contrast to English texts where we identified that simplification is related to a significant reduction of sentence lengths. Unlike in English texts, we observe significant differences between technical abstracts and PLSs in word lengths as measured by the number of syllables per word. Word lengths are significantly increased in PLSs. As a result of the increased word and sentence lengths in PLSs, FKGL is significantly higher in PLSs than in technical abstracts, which is contrary to the definition of FKGL and its interpretation for English texts.

POS-based metrics show that German technical abstracts use more nouns and numbers than PLSs, while PLSs use more function words. These differences are significant as indicated by noun, numbers, and function word ratios. We made the same observations also for English texts. Moreover, German technical abstracts have a significantly higher proportion of adjectives than their simplified counterparts. This is different from English texts where adjective ratio did not constitute a suitable metric to capture different grammatical characteristics of technical abstracts and PLSs. We make the opposite observation for the proportion of negations. While in English medical texts negation ratio is significantly higher in PLSs than in technical abstracts, the proportion of negations does not differ significantly between

Metric Type	Metric Name	Technical Abstract Mean	Plain Language Summary (PLS) Mean	p-value
Statistical	Words per sentence	15.899	16.898	0.048
	Syllables per word	2.107	2.197	**0.001**
	FKGL	15.472	16.924	**0.000**
POS-based	Noun ratio	0.353	0.302	**0.000**
	Adjective ratio	0.091	0.077	**0.000**
	Function word ratio	0.273	0.326	**0.000**
	Numbers ratio	0.078	0.023	**0.000**
	Negations ratio	0.003	0.004	0.047
Syntactic	Grammar frequency	0.885	0.965	**0.000**
	Average edit distance	0.759	0.761	0.721
	Dependency tree depth	6.127	6.663	**0.000**
	Noun phrase complexity	1.925	1.804	**0.000**
Semantic	Term frequency	0.004	0.005	**0.000**
	LMFM	0.379	0.269	**0.000**
Fluency	LM perplexity	277.119	196.558	0.211
	Average lexical chain length	0.006	0.010	**0.000**
	Average number of cross chains	0.750	0.395	**0.000**
	NSP score	0.855	0.875	0.075

Table 2: Means and p-values (p) obtained from paired sample t-tests for German technical abstracts and related plain language summaries (PLSs). Significant differences are highlighted in bold given a significance level of $p < 0.003$ after Bonferroni correction.

technical abstracts and PLSs in German medical texts. **Syntactic metrics** reveal that variability of sentence structures is significantly higher in German PLSs compared to German technical abstracts as indicated by grammar frequency, which is in contrast to English medical texts, where sentence structure variability does not seem to be an indicator of readability. However, no significant differences are observed for average edit distance. Unlike in English texts, where simplification is associated with a reduction of syntactic dependencies, dependency tree depth is even significantly higher in PLSs compared to technical abstracts in German medical texts. The length of composed noun phrases is significantly increased in German technical abstracts. This is similar to English medical texts.

Semantic metrics show that German PLSs use terms which are used more frequently in common language and which can be more easily predicted by German BERT MLM. These findings suggest that simplified PLSs are easier to understand on the semantic level than technical abstracts. The same observations were made for English medical texts.

Fluency metrics reveal no significant differences in terms of fluency at sentence level as indicated by LM perplexity. This is similar to English medical texts. Interestingly, LM perplexity on German texts is significantly higher compared to English texts indicating that the German LM is more uncertain in predicting the German texts than the corresponding English LM in predicting the English texts. This might be caused by the different properties of the underlying LMs arising from the composition of pre-training texts and their vocabularies. Significant differences for German texts are observed at full text level as indicated by average lexical chain length and average number of cross chains. As for English texts, lexical chains are significantly shorter in technical abstracts, while the number of intersecting chains is significantly higher in technical abstracts compared to PLSs. Coherence of adjacent sentences in German PLSs is slightly higher in PLSs than in technical abstracts as measured by NSP score. However, as in English texts, NSP does not differ significantly.

5 Discussion

Our study results show that two third out of 18 readability metrics sufficiently capture statistical, POS-based, syntactical, semantic, and fluency-related characteristics of English medical texts differentiating technical abstracts from PLSs. Similarly, 13 out of 18 readability metrics revealed significant differences between technical abstracts and PLSs for German medical texts. Our results show that semantic and fluency metrics can be directly transferred from English to German medical texts and can be interpreted in the same way. Contrastively, statistical, POS-based, and syntactic metrics behave differently on German texts compared to English texts and thus require careful interpretation. We assume that language-specific characteristics are the reason.

Comparing English with German medical texts, we recognize that sentences in German technical abstracts have about ten words less than sentences in English technical abstracts. Also after simplification, sentences in German texts are still more than four words shorter on average. Mukherjee et al., 2017 report on similar sentence lengths in technical Spanish medical texts as we observed for English technical abstracts. This indicates that sentences in German technical abstracts are characteristically short compared to other languages. As a result, there might be less potential to improve readability by reducing sentence lengths. Additionally, it is common in German technical texts to apply nominalization. During nominalization, verbs are transformed into nouns forming new noun phrases together with additional determiners and adjectives. Thus, nominalization results in an increase of noun phrases as indicated by higher noun ratios and noun phrase complexities and might also lead to an increase in the proportion of adjectives we have observed in German technical abstracts but not in English technical abstracts. Another variant is to combine nominalized nouns with prepositions, which can often replace whole subordinate clauses. This might result in shorter sentences as indicated by the average number of words per sentence and might reduce depths of syntactic dependencies as indicated by dependency tree depths.

Interestingly, we observed a high variability in terms of sentence structures in English technical abstracts, which remained nearly constant after simplification. Contrastively, in German technical abstracts, variability of sentence structures is lower compared to English technical abstracts. However, sentence structure variability significantly increased after simplification in German medical texts. On the one hand, this is surprising since English follows the subject-verb-object sentence structuring scheme while German has no fixed sentence structuring scheme and is thus more flexible in this regard. On the other hand, the lower sentence structure variability in German technical abstracts might be also related to the shorter and more compact sentences resulting from nominalizations which might reduce flexibility of how sentences can be structured. We provide examples of nominalizations found in German medical texts in Appendix A.

6 Conclusions

Research in MTS has mainly focused on English medical texts. To evaluate the suitability of available readability metrics to indicate readability of German medical texts, we evaluated 18 statistical, POS-based, syntactic, semantic and fluency metrics from related studies using paired sample t-tests on paired texts consisting of technical abstracts and related simplified summaries from English and German scientific articles, respectively. We found that semantic and fluency metrics suitable for indicating readability of English medical texts can be successfully applied to German medical texts. However, statistical metrics including sentence and word length-based metrics, POS-based metrics including the proportions of adjectives and negations, and syntactic metrics measuring complexity of syntactic dependencies and variability of sentence structures seem to be aligned with language-specific characteristics. We found that they behave differently and have to be interpreted carefully when they are applied to German medical texts. Further research is required to investigate language-specific characteristics of German medical texts and how these influence readability. Our results emphasize the need to further advance MTS research to other languages.

7 Limitations

We emphasize two limitations of our study. First, the datasets included in our study are limited in size, which might affect expressiveness of our results. The datasets originate from a single source and the PLSs follow a simplification guide

provided by the Cochrane editorial team, which could induce bias. Although the simplification guide is created by experts, the PLSs are written by the article authors themselves not by language professionals. We also admit that readability of medical texts might depend on the concrete topic covered. Here, we focused on texts related to cardiovascular diseases and cancer. However, medical texts discussing other topics might reveal different results. The non-English texts are translated by volunteering native speakers, professional translators, and Cochrane-internal translator teams using machine translation in combination with human post-editing. However, they are not written by the original authors, which might also induce bias. Second, our study focuses on evaluating readability metrics which enable to differentiate and characterize difficult and simplified medical texts. Although the PLSs, are intended to communicate evidences from medical studies to laypeople in a comprehensible way, our results do not allow to draw conclusions of how the readability of these texts is actually perceived by laypeople. This will be part of future work.

Acknowledgments

This work has been conducted at the Fraunhofer Institute for Digital Medicine MEVIS. This work has been supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) in the call “Innovative und praxisnahe Anwendungen und Datenräume im digitalen Ökosystem GAIA-X” under contract number 68GX21001C.

References

Mohammad Alarifi, Timothy Patrick, Abdulrahman Jabour, Min Wu, and Jake Luo. 2020. Full Radiology Report through Patient Web Portal: A Literature Review. *International Journal of Environmental Research and Public Health,* 17(10), 3673.

Mohammad Alarifi, Timothy Patrick, Abdulrahman Jabour, Min Wu, and Jake Luo. 2021. Understanding patient needs and gaps in radiology reports through online discussion forum analysis. *Insights into Imaging (2021) 12:50*, pages:1-9.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12), pages: 14093-14101.

Christina Baun, Marianne Vogsen, Marie Konge Nielsen, Poul Flemming Høilund-Carlsen, Malene Grubbe Hildebrandt. 2020. Perspective of Patients With Metastatic Breast Cancer on Electronic Access to Scan Results: Mixed Methods Study. *Journal of Medical Internet Research,* 22(2), e15723.

Bundesministerium für Gesundheit (BMG). 2024. Die elektronische Patientenakte (ePA). Bundesministerium für Gesundheit. https://www.bundesgesundheitsministerium.de/elektronische-patientenakte, accessed: 09/15/2024.

Joshua K. Cho, Hanna M. Zafar, and Tessa S. Cook. 2020. Use of an Online Crowdsourcing Platform to Assess Patient Comprehension of Radiology Reports and Colloquialisms. *American Journal of Roentgenology*, 214(6), pages: 1316-1320.

Judith Deppe. 2024. About translations at Cochrane. https://documentation.cochrane.org/display/TH/About+translation+at+Cochrane, accessed: 11/02/2024.

Koen Dercksen and Arjen P. de Vries. 2020. First Steps Towards Patient-Friendly Presentation of Dutch Radiology Reports.

Ashwin Devaraj, Iain J. Marshall, Byron C. Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages: 4972-4984.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pages: 4171-4186.

Andrew J. Gunn, Brian Gilcrease-Garcia, Mark D. Mangano, Dushyant V. Sahani, Giles W. Boland, and Garry Choy. 2017. JOURNAL CLUB: Structured Feedback From Patients on Actual Radiology Reports: A Novel Approach to Improve Reporting Practices. *American Journal of Roentgenology*, 208(6), pages: 1262-1270.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Inductrial-strenght Natural Language Processing in Python. 10.5281/zenodo.1212303

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stübner, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2024. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified

Radiology Reports. *European Radiology*, 34(5), pages: 2817-2825.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, Junyi Jessy Li. 2023. Multilingual Simplification of Medical Texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, pages:16662-16692.

David Kauchak, Gondy Leroy, and Alan Hogue. 2017. Measuring Text Difficulty Using Parse-Tree Frequency. *Journal of the Association for Information Science and Technology,* 68(9), pages: 2088-2100.

Alla Keselman, Laura Slaughter, Catherine Arnott Smith, Hyeoneui Kim, Guy Divita, Allen Browne, Christopher Tsai, Qing Zeng-Treitler. 2007. Towards Consumer-Friendly PHRs: Patients' Experience with Reviewing Their Health Records. In *AMIA Annual Symposium Proceedings*, pages: 399-403.

Alla Keselman and Catherine Arnott Smith. 2012. A classification of errors in lay comprehension of medical documents. *Journal of Biomedical Informatics* 45(6), pages: 1151-1163.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch.

Nicholas Kloehn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P Yuan, Debra Revere. 2018. Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish. *Journal of Medical Internet Research* 20(8), e10779.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Keep it Simple: Unsupervised Simplification of Multi-Paragraph Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, (Volume 1: Long Papers), Association for Computational Linguistics, pages: 6365-6378.

Ben Lambert. 2023. belambert/edit-distance : v1.0.6. https://github.com/belambert/edit-distance/tree/v1.0.6.

Gondy Leroy, Stephen Helmreich, and James R. Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, 79(6), pages: 438-449.

Gondy Leroy, James E. Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of Medical Internet Research* 15(7), e2569.

Qing Lyu, Josh Tan, Michael E. Zapadka, Janardhana Ponnatapura, Chuang Niu, , Kyle J. Myers, Ge Wang, and Christopher T. Whitlow. 2023. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1), 9.

Antonio Menta and Ana Garcia-Serrano. 2022. Controllable Sentence Simplification Using Transfer Learning. Proceedings of the Working Notes of CLEF.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, and Nils Hammerla. 2021. Towards more patient friendly clinical notes through language models and ontologies. In *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, pages: 881-890.

Partha Mukherjee, Gondy Leroy, David Kauchak, Brianda Armenta Navarrete, Damian Y. Diaz, and Sonia Colina. 2017. The Role of Surface, Semantic and Grammatical Features on Simplification of Spanish Medical Texts: A User Study. In *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, pages 1322-1331.

Partha Mukherjee, Gondy Leroy, and David Kauchak. 2019. Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study. *IEEE Journal of Biomedical and Health Informatics*, 23(5), pages: 2164-2173.

Babak Naderi, Salar Mohtaj, Karan Karan, Sebastian Möller. 2019. Automated Text Readability Assessment for German Language: A Quality of Experience Approach. *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, Berlin, Germany, pages:1-3.

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, Vijay Mago. 2022. Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning-Based Text Simplification Approach. *JMIR Medical Informatics,* 10(11), e38095.

Nicole Pitcher, Denise Mitchell, and Carolyn Hughes. 2022. Template and guidance for writing a Cochrane Plain language summary. Cochrane.

Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text Simplification Using Consumer Health Vocabulary to Generate Patient-Centered Radiology Reporting: Translation and Evaluation. *Journal of Medical Internet Research*. 19(12), e417.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Charlie Rogers, Sophie Willis, Steven Gillard, and Jane Chudleigh. 2023. Patient experience of imaging reports: A systematic literature review. *Ultrasound* 31(3), pages: 164-175.

Kristine Sørensen, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, Helmut Brand, and (HLS-EU) Consortium Health Literacy Project European. 2012. Health literacy and public health: A systematic review and integration of definition and models. *BMC public health*, 12, pages: 1-13.

Robyn Speer. 2022. rspeer/wordfreq: v3.0. 10.5281/zenodo.7199437

Stefan Schweter. 2020. German GPT-2 model. 10.5281/zenodo.4275046.

Johannes Valbjorn. 2023. sloev/spacy-syllables: v3.0.2. https://github.com/sloev/spacy-syllables.

Zarah Weiss, Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages: 141-153, Seattle, Washington. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, et al.. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38-45, Online. Association for Computational Linguistics.

Susan S Woods, Erin Schwartz, Anais Tuepker, Nancy A Press, Kim N Nazi, Carolyn L Turvey, and W Paul Nichol. Patient Experiences With Full Electronic Access to Health Records and Clinical Notes Through the My HealtheVet Personal Health Record Pilot: Qualitative Study. 2013. Journal of Medical Internet Research, 15(3), e65.

Ziyu Yang, Santhosch Cherian, and Slobodan Vucetic. 2023. Data Augmentation for Radiology Report Simplification. In *Findings of the Association for Computational Linguistics: EACL 2023*, Association for Computational Linguistics, pages: 1922-1932.

Qing Zeng-Treitler, Hyeoneui Kim, Sergey Goryachev, Alla Keselmann, Laura Slaughter, Catherine Smith. 2007. Text Characteristics of Clinical Reports and Their Implications on the Readability of Personal Health Records. In *MEDINFO 2007*, IOS Press, pages: 1117-1121.

Richard Zowalla, Daniel Pfeifer, Thomas Wetter. 2023. Readability and topics of the German Health Web: Exploratory study and text analysis. PLoS ONE 18(2):e0281582

A Example Simplifications of German Medical Texts

We extracted original sentences from German technical abstracts and related simplified sentences from the corresponding PLSs to exemplarily show the particularities of German language which affect readability. Examples are shown in Table 3.

B Metric Implementation Details

Text preprocessing is performed using the open-source NLP library *spaCy* (Honnibal et al., 2020) and *Spacy Syllables* (Valbjorn, 2023) with the programming language Python. We use *spaCy* version 3.6.1 and *Spacy Syllables* version 3.0.2. For preprocessing the English dataset, we use the *spaCy* NLP pipeline *en_core_web_lg* in version 3.6.0. For preprocessing the German dataset, we use the *spaCy* NLP pipeline *de_core_news_lg* in version 3.6.0.

The POS-based readability metrics are calculated according to Equation 4. However, the set of target POS categories is adapted with respect to the respective metric. The set of applied target POS categories for each metric is shown in Table 4. The pipeline-specific label schemes are documented in the *spaCy* documentation for the English[3] and German[4] pipelines. For calculating noun phrase complexity, noun phrases are obtained from the *spaCy* "*noun_chunks*" property.

To calculate Levenshtein distances, which is required to calculate average edit distance, the Python package *edit-distance* in version 1.0.6 (Lambert, 2023) is used. To calculate term frequency, we use the Python library *wordfreq* in

Technical Abstract	**Plain Language Summary (PLS)**
Beurteilung der Wirksamkeit von Interventionen […] (*noun phrase with base noun „Beurteilung" and „der Wirksamkeit von Interventionen" as genitive object*) Assessing the effectiveness of interventions […]	Wir wollten herausfinden, ob eine Behandlung wirkt […] (*subordinate clause resolving „Beurteilung der Wirksamkeit von Interventionen"*) We wanted to find out whether a treatment works […]
Allerdings können aus der derzeitigen Evidenz keine Schlussfolgerungen bezüglich eines optimalen systolischen Blutdruckziels (*noun phrase with base noun „Blutdruckziel" functioning as a n adverbial clause*) […] gezogen werden. However, from the current evidence no conclusions can be drawn regarding an optimal systolic blood pressure […]	Außerdem gibt es nicht genug Evidenz um festzustellen, welches Blutdruckziel […] das beste ist. (*subordinate clause resolving: „bezüglich eines optimalen systolischen Blutdruckziels"*) Furthermore, there is not enough evidence to determine which blood pressure target is best.
Das primäre Ziel war, die Wirksamkeit von […] Interventionen zur Behandlung von Patienten mit Schlaganfall und Angststörungen oder -symptomen zu untersuchen. (*noun phrases with base nouns „Wirksamkeit" and „Behandlung"*) The primary aim was to investigate the effectiveness of [...] interventions for the treatment of patients with stroke and anxiety disorders or symptoms.	Ziel war herauszufinden, ob es Behandlungen gibt, die Angstsymptome reduzieren und […] die Lebensqualität […] verbessern können. (*subordinate clauses resolving: „die Wirksamkeit von Interventionen zur Behandlung von Patienten mit Schlaganfall und Angststörungen oder -symptomen zu untersuchen"*) The aim was to find out whether there are treatments that can reduce anxiety symptoms and [...] improve quality of life [...]

Table 3: Example sentences from German technical abstracts and related sentences from corresponding simplified PLSs showing particularities of German language, which affect readability. We underlined corresponding sentence sections in technical abstracts and PLSs and added annotations to clarify the sentence structure.

[3] https://spacy.io/models/en

[4] https://spacy.io/models/de

Metric	Target POS Categories	
	English	**German**
Noun Ratio	NOUN, PROPN (Universal POS tags)	NOUN, PROPN (Universal POS tags)
Adjective Ratio	ADJ (Universal POS tags)	ADJ (Universal POS tags)
Function Word Ratio	AUX, DET, ADP (Universal POS tags) MD, WDT, WP, WP$, WRB (Pipeline-specific POS tags)	AUX, DET, ADP (Universal POS tags) VMFIN, VMINF, VMPP, PWS, PWAT, PWAV (Pipeline-specific POS tags)
Numbers Ratio	NUM (Universal POS tags)	NUM (Universal POS tags)
Negations Ratio	neg (Pipeline-specific dependency labels)	ng (Pipeline-specific dependency labels)

Table 4: Sets of target part-of-speech (POS) categories and dependency labels used for calculating POS-based readability metrics.

version 3.1.1. We obtain all LMs used to calculate LMFM, LM perplexity, and NSP score from the *Hugging Face Hub*[5]. We calculate metrics locally using the *Transformers* (Wolf et al., 2020) library in version 4.28.1.

[5] https://huggingface.co/models