MEDREADME: A Systematic Study for Fine-grained Sentence Readability in Medical Domain

Anonymous EMNLP submission

Abstract

Medical texts are notoriously challenging to read. Properly measuring their readability is the first step towards making them more accessible. Here, we present the first systematic study on fine-grained readability measurements at both sentence- and span-level in the medical domain. We first introduce a new dataset MEDREADME, which consists of manually annotated readability ratings for 1,924 sentences, and fine-grained complex span annotation on 4,520 sentences, featuring two novel "Google-Easy" and "Google-Hard" categories. It supports our investigation into the quality of 15 widely used medical text simplification datasets. Enabled by our high-quality corpus, we benchmark several state-of-the-art sentence-level readability metrics, including unsupervised, supervised, and prompting-based methods using recently developed large language models (LLMs). Informed by our fine-grained complex span annotation, we find that adding a single feature, capturing the number of jargon spans, into existing sentence readability formulas can significantly improve their correlation with human judgments, and also make them more stable. We will publicly release our dataset and code.

1 Introduction

> *If you can't measure it, you can't improve it.*
>
> – Peter Drucker

Timely disseminating reliable medical knowledge to those in need is crucial for public health management (August et al., 2023). Trustworthy platforms like Merck Manuals and medical Wikipedia contain extensive information, and research papers introduce the latest findings, including emerging medical conditions and treatments (Joseph et al., 2023). However, comprehending these resources can be a challenging task due to their technical nature and the extensive use of specialized terminology (Zeng et al., 2005). As the first step to making them more accessible, properly measuring the readability of medical texts is crucial (Rooney et al., 2021; Echuri et al., 2022). However, a high-quality dataset is lacking for evaluating sentence readability metrics in the medical domain and further improving them by studying finer-grained span-level information.

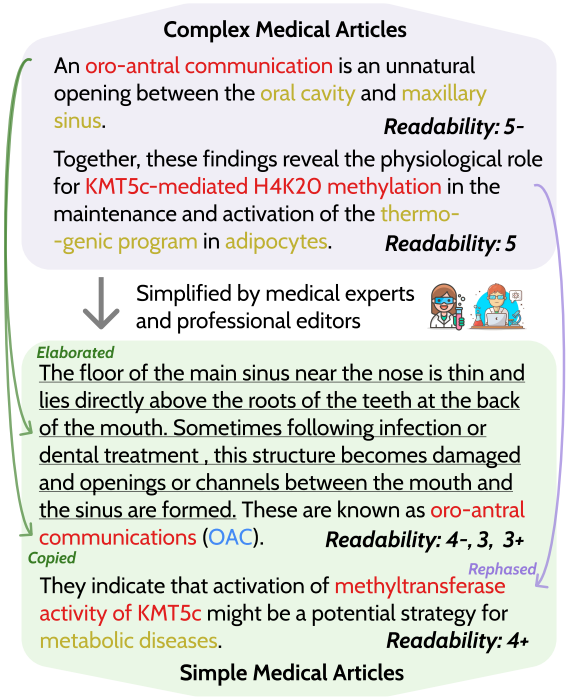


Figure 1: An illustration of our dataset, with sentence readability ratings for 1,924 sentences and fine-grained complex span annotation on 4,520 sentences, including "Google-Hard" and "Google-Easy", abbreviations, and general complex terms, etc. We also analyze how medical jargon are being handled during simplification. e.g., a Google-Hard "oro-antral communication" is copied and elaborated. Some jargon are ignored for clarity.

In this work, we present a systematic study for sentence readability in medical domain, including a new dataset, an analysis of existing corpora, improvements to the state-of-the-art metrics, and trained models. As demonstrated in Figure 1, we first construct MEDREADME, consisting of 1,924 readability ratings at sentence-level, in addition to fine-grained complex span annotation on 4,520 sen-

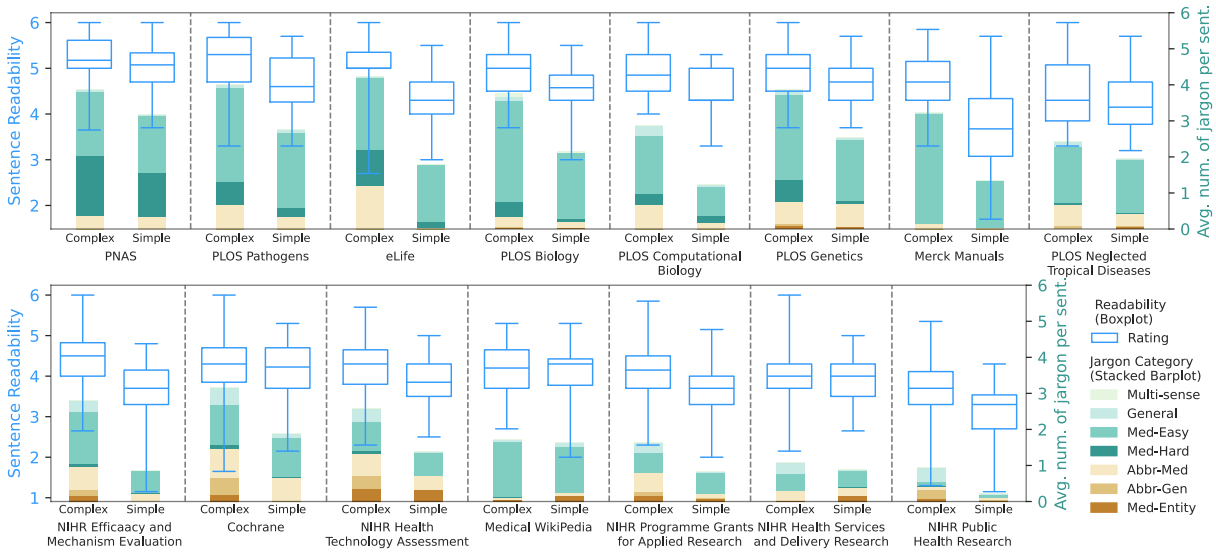


Figure 2: The distribution of average sentence readability (boxplot on the left y-axis) and the average number of jargon spans per category (stacked barplot on the right y-axis) in each sentence across both "complex" and "simplied" versions for 15 commonly used text simplification resources. Sentences with higher readability scores require a higher level of education to comprehend. The readability of sentences in different resources varies significantly.

tences (§3). It covers 15 resources that have been widely used in medical text simplification research. The readability ratings are annotated using a rank-and-rate interface (Maddela et al., 2023) based on the CEFR scale (Arase et al., 2022), which is shown to be more reliable than the other methods (Naous et al., 2023). To capture the reasons behind the sentence-level rating, we ask annotators to highlight any span that they find hard to understand, and categorize the reason using a 7-class taxonomy that we developed. Considering that "the majority of people seek health information online began at a search engine",[1] we introduce two "Google-Easy" and "Google-Hard" classes, providing a fresh perspective beyond binary or 5-point Likert scales.

Our dataset addresses three limitations in prior work. First, previous research on sentence readability use document-level ratings as approximation, which is shown to be inaccurate (Arase et al., 2022; Cripwell et al., 2023). Furthermore, recent work with sentence-level ratings mainly covers data from general domains, such as Wikipedia (De Clercq and Hoste, 2016), news (Štajner et al., 2017; Brunato et al., 2018), and textbooks for ESL learners (Arase et al., 2022), which are very different from specialized fields, such as medicine (Choi and Pak, 2007). Lastly, existing work separate the research on sentence readability and complex jargon terms, therefore missing the possible correlations between them (Kwon et al., 2022; Naous et al., 2023).

Enabled by our high-quality annotated corpus, we find that there are significant variations in the readability and jargon distribution across 15 resources (Figure 2). In addition, some of their topics may not be directly relevant to laypersons. Hence, we argue that the current research practice of simply combining all available data for training should be re-considered (more details in §4.1).

In the evaluation of sentence readability metrics, we find that unsupervised methods based on lexical features do not perform very well, whereas RSRS score (Martinc et al., 2021) using a weighted sum of words' log probability achieves better performance. Prompting large language models such as GPT4 and Llama2 (Touvron et al., 2023) with 5-shot shows strong performance, yet outperformed by fine-tuned models in a much smaller size (§4.2).

Based on our span-level annotation, we find that medical jargon significantly impact readability, even more so than the average sentence or word length (§4.3). We add a single feature that captures the "number of jargon" in a sentence into existing readability formulas, and find it can significantly improve their correlations with human judgments as well as stability. To identify the complex span in a real scenario, we train a complex span identification model using our annotations (§5). Using our best performing model, the proposed method achieves a 0.714 Pearson correlation for supervised metrics and 0.792 for the unsupervised ones (§4.4).

2 Related Work

Readability Measurement in Medical Domain. Unsupervised metrics, such as FKGL (Kincaid et al., 1975), ARI (Smith and Senter, 1967), SMOG

[1] https://tinyurl.com/seek-health-info-online

(Mc Laughlin, 1969), and Coleman-Liau index (Coleman and Liau, 1975) do not need training data and have been widely adopted in existing research on the medical readability analysis (Fu et al., 2016; Chhabra et al., 2018; Xu et al., 2019; Devaraj et al., 2021a; Kruse et al., 2021; Guo et al., 2022; Kaya and Görmez, 2022; Hartnett et al., 2023, *inter alia*). However, their reliability has been questioned (Wilson, 2009; Jindal and MacDermid, 2017; Devaraj et al., 2021b), as they mainly rely on the combination of lexical features, such as sentence length, word length, and syllable count. RSRS score (Martinc et al., 2021) utilizes the log probability of words from a pre-trained language model such as BERT (Devlin et al., 2019), and supervised metrics formulate the problem as a regression or classification task and fine-tune a machine learning model on the annotated corpora (Arase et al., 2022; Naous et al., 2023), whereas their performance is unclear for the medical field. Enabled by our high-quality dataset, we comprehensively benchmark state-of-the-art metrics in the medical domain (§4.2), and also further improve their performance (§4.4).

Complex Span Identification in Medical Domain. There are two prior studies covering the data from medical domain. CompLex 2.0 (Shardlow et al., 2020) consists of sentences from Bible, Europarl, and biomedical texts. Given one span in a sentence, annotators rate the difficulty using a 5-point Likert scale. However, it only annotates spans with one or two tokens, and the latter are selected based on specific syntax patterns, such as adjective-noun or noun-noun. MedJEx corpus (Kwon et al., 2022) have 95k binary jargon annotation for 18k sentences from the medical notes, whereas the dataset is licensed. Other work mainly focuses on the general domains such as news and Wikipedia, including CW corpus in SemEval 2016 shared task (Shardlow, 2013; Paetzold and Specia, 2016) and CWIG3G2 corpus in SemEval 2018 (Yimam et al., 2017, 2018). Our data is based on open-access resources, and contains complex span annotation in a finer-grained 7-class on 4,520 sentences. All the complex spans in each sentence are identified by annotators without any restriction.

3 The MEDREADME Corpus

This section presents the detailed procedure for constructing the Medical Readability Measurement (MEDREADME) corpus. Our corpus consists of 180 complex-simple article pairs randomly sampled from 15 resources. In total, the data has 4,520 sentences, and is split into train/dev/test sets with 2,587/784/1,140 sentences (§3.1). To reliably evaluate existing readability metrics, we first collect readability judgements for 1,924 sentences in the dev and test sets (§3.2). Based on our pilot study and feedback from annotators, we find that specialized terminology seriously affects the readability level of texts from professional domains, such as medicine. Therefore, we ask annotators to identify any span they deem as complex and categorize the reason for all 4,520 sentences (§3.3), which are then used to train and evaluate our complex span identification model (§5). Finally, we conduct an initial data-driven analysis to show how the medical jargon are being handled by professional experts during the manual simplification process (§3.4).

Source of the Publication	Avg. #Sent. Comp./Simp.	Avg. Sent. Len. Comp./Simp.
Public Library of Science (PLOS)		
Biology	8.3 / 8.2	28.2 / 26.8
Genetics	10.2 / 6.2	28.9 / 30.3
Pathogens	8.9 / 7.2	30.7 / 29.5
Computational Biology	9.1 / 7.2	29.3 / 27.4
Neglected Tropical Diseases	10.2 / 8.0	29.3 / 26.4
National Institute for Health and Care Research (NIHR)		
Public Health Research	23.4 / 14.3	26.2 / 20.5
Health Technology Assessment	25.1 / 12.9	27.3 / 25.7
Efficacy and Mechanism Evaluation	22.6 / 14.9	28.2 / 21.4
Programme Grants for Applied Research	27.6 / 14.2	27.6 / 22.6
Health Services and Delivery Research	23.2 / 14.1	27.9 / 23.2
Medical Wikipedia	5.4 / 5.8	23.3 / 19.4
Merck Manuals (medical references)	5.0 / 5.6	23.8 / 16.3
eLife (biomedicine and life sciences)	6.5 / 15.6	27.0 / 26.3
Cochrane Database of Systematic Reviews	25.4 / 16.1	27.3 / 22.2
Proc. of National Academy of Sciences	9.1 / 5.5	27.2 / 24.1

Table 1: Average # of sentences and their length for 180 sampled parallel articles (segments) from 15 resources.

3.1 Data Collection and Preprocessing

The 15 data resources that we considered include the abstract sections and plain language summaries from scientific papers, such as National Institute for Health and Care Research (NIHR) and "the highest standard in evidence-based healthcare" Cochrane Review,[2] for which we use the aligned article pairs from prior studies (Devaraj et al., 2021a; Goldsack et al., 2022; Guo et al., 2022). We also include segment and paragraph pairs for the parallel versions of medical references from trusted online platforms, such as Merck Manuals[3] and Medical Wikipedia. The detailed introduction of each resource is pro-

[2] https://www.cochranelibrary.com/
[3] https://www.merckmanuals.com/

Category	Definition	Example	Tok. Len.	%
Medical Jargon			2.2±1.5	68.6%
Google-Easy	Medical terms that can be easily understood after a quick search.	Schistosoma mansoni is a parasitic infection common in the tropics and sub-tropics.	2.0±1.2	56.9%
Google-Hard	Medical terms that require extensive research before a layperson can possibly understand them.	. . . retains limited DNA-processing activity, albeit via a distributive binding mechanism.	3.2±2.5	7.5%
Name Entity	Brand or organization name, excluding general medical terms such as drugs and equipments.	While vaccination with BioNTech and Moderna mostly causes only mild and typical . . .	2.7±2.2	4.1%
General Complex	Terms that are outside the vocabulary of 10-12th graders and not specific to the medical domain.	Treatments used to ameliorate symptoms and reduce morbidity include opiates, sedatives . . .	1.9±1.2	10.2%
Multi-sense	Spans that have different meanings in the medical context compared to their general use.	. . . in structural and/or functional aspects of the interaction with the insect vector.	1.0±0.1	0.5%
Abbreviation			1.1±0.4	20.8%
Medical Domain	Abbreviations that have a specific meaning in the medical domain.	. . . 4,433 were alive and not withdrawn at an LTFU participating center.	1.1±0.4	16.6%
General Domain	Abbreviations that belong to the general domain.	. . . as low risk of bias (95% CI 0.37 to 1.53).	1.0±0.2	4.2%

Table 2: A taxonomy ($\mathcal{I}$) of complex spans in medical domain. In each example, the complex spans are marked with a red background. The "medical jargon" and "abbreviation" rows are based on the aggregation of sub-categories.

vided in Appendix B. Table 1 presents the basic statistics of 180 sampled article (segment) pairs. We use the Stanford CoreNLP toolkit (Manning et al., 2014) to perform sentence splitting and word tokenization. We also further improve it by adding additional rules to fix splitting errors caused by the diverse formatting of medical text, and manually correct any errors before further annotation.

3.2 Sentence-level Readability Annotation

To collect ground-truth judgements, we hire three in-house undergrads with rich experience to double annotate the readability ratings for 1,924 sentences. A two-hour training session is provided, during which annotators are asked to annotate 150 sentences and discuss untill consensus. Our annotating setup utilizes the "rank-and-rate" interface (Naous et al., 2023) and the CEFR scale (Arase et al., 2022), with several improvements.

CEFR Scale. Following prior work (Arase et al., 2022), we use the Common European Framework of Reference for Languages (CEFR), which is the most widely used international criteria to define learners' language proficiency. CEFR assesses language skills by a 6-level scale with a detailed guideline,[4] from beginners (A1) to advanced mastery (C2), which are denoted as level 1 (easiest) to level 6 (hardest) in our interface. A sentence's readability is determined based on the CEFR level at which an individual can understand it with ease, thus providing a standardized measure. It has been shown to be more reliable compared to 5-point (Štajner et al., 2017), 7-point (Brunato et al., 2018), and 0-100 scale (De Clercq and Hoste, 2016), which are based on annotators' subjective judgments. As medical texts are naturally on the harder side, we introduce the use of "+" and "-" signs to differentiate the nuance in readability, e.g., "3+" and "3-", in addition to the integer level. They are treated as 3.3 and 2.7 when converting to numeric scores.

Rank-and-Rate Framework. Six sentences are shown on each page, and annotators are instructed to rank them from most to least readable first, and then rate each sentence using the 6-point CEFR standard. The interface is shown in Appendix F. Compared to rating each sentence individually, this method enables annotators to compare and contrast sentences within each batch, leading to higher annotator agreement (Maddela et al., 2023) and more engaging user experience (Naous et al., 2023).

Quality Control. For every new sentence from the MEDREADME corpus, we sample another sentence with comparable length from README++ dataset (Naous et al., 2023) as a "contrasting sentence". Therefore, each page consists of three new sentences and three contrasting ones whose ratings are known. All annotators are compensated at $18 per hour based on school standards, well above the minimum wage. Their performance is constantly monitored using the contrasting sentences, and they are asked to spend at least three minutes on every page. We double annotate the 1,924 sentences and merge the scores by average. The inter-annotator

[4] https://tinyurl.com/CEFR-Standard/

Sources	Length	FKGL	ARI	SMOG	RSRS	FKGL-Jar	ARI-Jar	SMOG-Jar	RSRS-Jar
Cochrane	0.628	0.743	0.689	0.749	0.826	0.717	0.719	0.726	0.721
PNAS	0.554	0.480	0.441	0.615	0.594	0.660	0.650	0.685	0.657
NIHR Series	0.529	0.482	0.455	0.661	0.659	0.577	0.583	0.632	0.616
eLife	0.505	0.196	0.244	0.371	0.467	0.644	0.638	0.690	0.733
PLOS Series	0.436	0.414	0.413	0.446	0.613	0.716	0.717	0.704	0.707
Wiki	0.352	0.400	0.368	0.471	0.670	0.677	0.681	0.785	0.703
MSD	0.259	0.618	0.576	0.604	0.694	0.836	0.835	0.805	0.859
Mean ± Std	0.466 ± 0.127	0.476 ± 0.173	0.455 ± 0.143	0.56 ± 0.134	0.646 ± 0.109	0.690 ± 0.080	0.689 ± 0.080	0.718 ± 0.060	0.714 ± 0.076

Table 3: Pearson correlation between human ground-truth readability and each **unsupervised** readability metric. NIHR and PLOS are aggregations of 5 sources for each. All correlations are statistically significant. "**-Jar**" denotes incorporating predicted jargon into the readability formula. We find that incorporating a "number-of-jargon" feature effectively improves the correlation over existing metrics, as demonstrated by the average correlation.

agreement is 0.742 measured by Krippendorff's alpha (Krippendorff, 2011). On the contrasting sentences, our annotation achieves a Pearson correlation of 0.771 with the original ratings from README++, demonstrating our quality.

3.3 Fine-trained Complex Span Annotation

We propose a new taxonomy to comprehensively capture the categories of complex spans that appeared in the medical texts, as shown in Table 2.

"Google-Hard" Jargon. In pilot study, we find that some medical terms, such as "Tiotropium bromide" (*a drug*) and "Plasmodium" (*an insect*), can be grasped after a quick Google search, although they are outside the vocabulary of many people. Some other phrases, such as "anti-tumour necrosis factor failure" and "processive nucleases", will require extensive research before a layperson can possibly (or still not) understand them, even though some of them contain short or common words. This motivates us to propose a unique category "Google-Hard" medical jargon, which is separate from jargon that are "Google-Easy" or "Name-Entity".

Annotation Agreement. After receiving a two-hour training session, two undergraduate annotators independently annotate each of the 4,520 sentences using a web-based annotation tool, BRAT (Stenetorp et al., 2012). Annotators are paid $18 per hour. An adjudicator then further inspects the annotation and discusses any significant disagreements before finalizing the annotation. The inter-annotator agreement is 0.631 before adjudication, measured by token-level Cohen's Kappa (Cohen, 1960). The annotation interface is in Appendix G.

3.4 How Jargon are being Handled?

To study how jargon are handled during the manual simplification process, we randomly sample 200 jargon and manually analyze the operation applied to them. The results are presented in Table 4. We find that the majority part of jargon in both categories got deleted. Compared to Google-Easy, "Google-Hard" jargon got copied less, and are being rephrased and explained more often. This findings indicate that trained editors adopt different strategies to handle jargon in different complexity.

Operation	Google-Easy	Google-Hard
Kept	22%	13% (↓ 9%)
Deleted	56%	52% (↓ 4%)
Rephrased	3%	10% (↑ 9%)
Kept + Explained	8%	8% (−)
Del.+ Explained	11%	17% (↑ 6%)

Table 4: The distribution of operations to 200 medical jargon (100 in each type), based on our manual analysis.

4 Readability in the Medical Domain

We present an in-depth study of sentence-level readability in the medical domain, including an analysis of existing dataset, an evaluation on SOTA metrics, and improvements to them. We find that the readability level of simplified texts varies significantly between 15 text simplification resources, as well as the relevance of the topics (§4.1). We then evaluate the performance of automatic readability metrics (§4.2), and find the consideration of jargon is important for accurately measuring sentence readability (§4.3). By adding a new feature to existing metrics, we significantly improve their correlation with human judgments and make them more stable (§4.4).

4.1 Variance of Readability across Resources

In Figure 2, we plot the distribution of sentence readability and numbers of jargon in each category for 15 text simplification resources. The complex versions are the original articles or abstract sections, and their simplified versions are written by the original authors or trained editors. Within each source, the simplified texts are rated as easier to understand than their complex counterparts, though the

Sources	5-shots			Trained on Each Corpus			The Trained + an Jargon Term		
	GPT4	GPT3.5	Llama 2-7b	ReadMe++	CEFR-SP	CompDS	ReadMe++$_{\text{Jar}}$	CEFR-SP$_{\text{Jar}}$	CompDS$_{\text{Jar}}$
Cochrane	0.908	0.460	0.549	0.858	0.899	0.870	0.849	0.841	0.783
PNAS	0.780	0.496	0.574	0.852	0.820	0.791	0.824	0.780	0.743
NIHR Series	0.713	0.553	0.580	0.824	0.753	0.706	0.686	0.695	0.633
eLife	0.538	0.317	0.127	0.594	0.715	0.608	0.803	0.811	0.776
PLOS Series	0.672	0.556	0.309	0.680	0.691	0.635	0.843	0.787	0.745
Wiki	0.670	0.527	0.429	0.824	0.709	0.607	0.620	0.711	0.673
MSD	0.766	0.376	0.328	0.784	0.778	0.757	0.881	0.917	0.862
Mean ± Std	0.721 ± 0.115	0.469 ± 0.092	0.414 ± 0.17	0.774 ± 0.1	0.766 ± 0.073	0.711 ± 0.101	0.787 ± 0.096	0.792 ± 0.076	0.745 ± 0.075

Table 5: Pearson correlation between human ground-truth readability and each **prompting** and **supervised** readability metric. NIHR and PLOS are aggregations of 5 sources for each. All numbers are averaged over five runs, and all correlations are statistically significant. denotes RoBERTa-large models. "**-Jar**" means adding a "jargon" term (more details in §4.4). Prompt-based methods are competitive, while still outperformed by fine-tuned models.

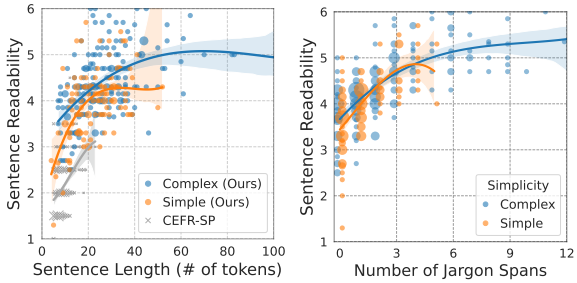


Figure 3: *Left*: Readability of sentences with different lengths. Compared to the CEFR-SP dataset (Arase et al., 2022), our corpus contains much longer sentences. *Right*: Readability of sentences with different numbers of jargon. The circle's radius reflects the number of overlapping points at each coordinate. We slightly shifted the points horizontally (±0.1) to improve clarity.

extent varies. However, compared across venues, some simplified texts are more challenging to read than the complex texts from other sources, suggesting that not all plain texts are "equally" simplified. In addition, some resources, such as "PLOS pathogens", are especially challenging for laypersons without domain-specific knowledge to understand. Finally, we also find the number and type of jargon used in each resource differ a lot.

The current research practice treats medical text simplification as an umbrella term, often concatenating all available corpora into a giant training set. However, we argue for a more cautious approach. For some resources, the "simplified" version remains quite complex, and the topics may not be directly relevant to laypersons. Therefore, the decision to include a corpus or not should be made after considering the intended audiences' desired simplification level and the application domain.

4.2 Evaluating Existing Readability Metrics

Enabled by our annotated corpus, we first evaluate commonly used sentence readability metrics.

Unsupervised Metrics. The Pearson correlations between ground-truth readability and each unsupervised metric are presented in the left half of Table 3, and their detailed formulations are provided in Appendix A. We also add sentence length as a baseline. We find that they generally do not perform very well. The language model-based RSRS score significantly outperforms the traditional feature-based metrics, among which SMOG performs best.

Type	#Spans	#Tokens	%Tokens
Medical Jargon	0.611	0.577	0.413
Abbreviation	0.306	0.300	0.203
General Complex	0.111	0.093	0.006
Multi-sense	0.040	0.040	0.024
All Categories	0.650	0.618	0.563

Table 6: The importance of each possible feature, measured by the Pearson correlation with ground-truth sentence readability on all corpora. For reference, the correlation between readability and sentence length is 0.470; the correlation with average word length is 0.143.

Supervised and Prompt-based Methods. The results are presented in Table 5. For the supervised methods, we fine-tune language models on existing corpora (Naous et al., 2023; Arase et al., 2022; Brunato et al., 2018) to predict the sentence readability. We also evaluate the performance of in-context learning by prompting large language models such as GPT4 and Llama 2 (Touvron et al., 2023) using 5-shot. The prompts are constructed following (Naous et al., 2023). More details and the full template can be found in Appendix E. We find that prompt-based methods achieve competitive results, e.g., GPT4 outperforms the strongest unsupervised metric RSRS, although they still fall behind supervised ones. In addition, models trained on ReadMe++ and CEFR-SP perform better because they use the same CEFR standard for annotation.

4.3 Jargon Drive Up Reading Difficulties

Based on our pilot study and feedback from annotators, we find that the specialized terminology,

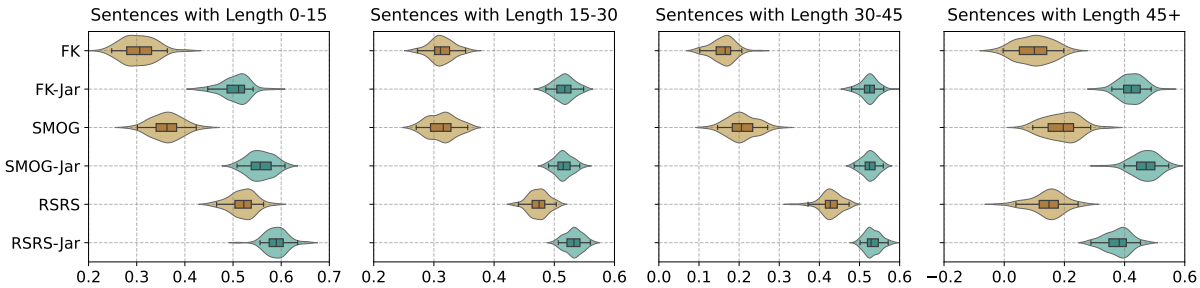


Figure 4: The 95% confidence intervals for Kendall Tau-like correlation (τ_{all}) between ground-truth readability annotation and predicted outputs from each automatic metric for sentences with different lengths, calculated by bootstrapping (Deutsch et al., 2021). In addition to a higher correlation with human judgments, incorporating jargon ("-Jar") makes each metric more stable, as shown by the smaller intervals.

while allowing for precise and concise communication among experts, significantly affects the difficulty level of texts in specialized domains, such as medicine. Failing to understand jargon prevents readers from getting specific details and even hinders them from parsing the overall structure of a sentence. The right part of Figure 3 presents the correlation between readability and the number of jargon spans, where we observe a generally positive correlation.[5] The correlation between readability and sentence length is plotted in left side of Fig. 3.

Based on our span-level annotation, we explicitly look into each type of fine-grained complex spans and measure their effects on the readability. Specifically, for each category, we design three features and compute their correlation with the sentence-level readability ratings, and also consider the collective impact for all jargon types. As shown in Table 6, we find that medical jargon significantly affects readability, and abbreviations follow in influence. The general complex terms and medical entities exhibit less effect. This nuanced understanding of how different types of jargon impact text comprehension can guide professionals to more precisely improve the accessibility of medical texts.

Of the three features we designed, "number-of-jargon-spans" and "number-of-jargon-tokens" show strong effectiveness and outperform the "percentage-of-jargon-tokens". We think this is likely because the first two features implicitly capture the sentence length to some extend, while the "percentage" feature cancels out the length factor.

4.4 Improve Existing Metrics by "-Jar"

To incorporate the consideration of jargon into existing metrics, we tune a weight α for the feature "number-of-jargon" for each metric, and add it to the current formula, as shown below:

$$\text{FKGL-Jar} = \text{FKGL} + \alpha \times \#\text{Jargon}$$

Here, "FKGL-Jar" denotes adding jargon into the FKGL score, similarly for other metrics with a suffix "-Jar". The weight α for each metric is tuned by grid search on the devset using gold annotation. As RSRS scores are smaller than 1, we scale them by timing 100 before doing parameter search. The right part in Table 3 and 5 reports the performance of each unsupervised and supervised method on the testset, after adding our proposed term. To simulate the real-word scenario, we use jargon predicted by our best performing complex span identification model (more details in §5), instead of the ground-truth annotation. The optimal weights we tuned for "FKGL-Jar", "ARI-Jar", "SMOG-Jar", and "RSRS-Jar" are 4.85, 6.43, 1.1, and 0.45, respectively. We find that introducing a single term significantly improves the correlation with human judgments.

To analyze the impact on sentences of different lengths, we conducted a length-controlled experiment. Figure 4 shows the 95% confidence intervals for the Kendall Tau-like correlation (τ_{all}) (Noether, 1981) between the ground-truth readability and predictions from each metric (Maddela et al., 2023). We find the proposed term is helpful for sentences at all lengths and is more significant for feature-based methods, such as SMOG. In addition, the incorporation of jargon makes the metrics more stable, as demonstrated by the narrower interval.

5 Fine-grained Complex Span Identification

Based on our analysis in §4.4, identifying complex spans in a sentence can help the judgment of its readability. It can also improve the performance of downstream text simplification system (Shardlow, 2014). We formulate this task as a NER-style sequential labeling problem (Gooding and Kochmar,

[5]The readability ratings are capped at 6.

Models	Token-Level			Entity-Level		
	Binary	3-Cls.	7-Cate.	Binary	3-Cls.	7-Cate.
Large-size Models						
BERT (2019)	86.1	80.9	67.9	78.5	74.1	43.9
RoBERTa (2019)	**86.8**	**82.3**	**68.6**	**80.2**	**75.9**	**67.9**
BioBERT (2020)	85.3	80.7	67.0	78.4	72.6	64.9
PubMedBERT (2021)	85.7	**82.3**	68.3	79.0	75.2	66.5
Base-size Models						
BERT (2019)	85.4	80.4	66.3	77.0	72.5	63.3
RoBERTa (2019)	86.2	81.7	68.0	79.7	75.2	66.6
BioBERT (2020)	84.2	79.6	66.4	77.1	72.8	64.1
PubMedBERT (2021)	85.2	81.2	67.7	78.5	74.8	66.3

Table 7: **Micro F1** of different systems for complex span identification on the MEDREADME testset. The **best** and second best scores within each model size are highlighted. Models are trained with fine-grained labels in seven categories and evaluated at different granularity.

2019), and utilize our annotated dataset to train and evaluate several models. We also study the performance of transfer learning by training models using existing datasets (Yimam et al., 2017; Paetzold and Specia, 2016), which are from general domains. The experimental results demonstrate the necessity for our medical-specific corpus.

Data and Models. The 4,520 sentences in our corpus is split into 2,587/784/1,140 for train, dev, and test sets. We mainly consider BERT/RoBERTa-based standard tagging models, initialized with different pre-trained embeddings. The implementation details are provided in Appendix C.

Evaluation Metrics. Our system will mainly be utilized to aid the judgment of sentence-level readability, where we found the numbers of jargon and jargon tokens are most helpful (§4.3). Serving this purpose, the following two metrics are mainly considered: (1) entity-level partial match, indicating the number of jargon, where the type of the predicted entity matches the gold entity and the predicted boundary overlaps with the gold one (Tabassum et al., 2020);[6] (2) token-level match, measuring the number of jargon tokens. For each metric, we conduct evaluations at three levels of granularity: (1) fine-grained level with seven categories, (2) associated higher-level classes (medical / general+multisense / abbreviation), and (3) binary judgements between jargon or not-jargon.

[6]We use the evaluation script released by the author at https://github.com/jeniyat/WNUT_2020_NER/tree/master/code/eval. We also report the exact match performance at entity-level in the Appendix D.

Training Corpus	Domain	# Sent.	Token	Entity
SemEval2016 (2016)	Wikipedia	200	38.6	29.0
CWIG3G2 (2017)	News, Wiki	1,988	46.4	28.7
MEDREADME (Ours)	Medical Articles	4,520	86.8	80.2

Table 8: F1 on the testset of MEDREADME for models trained on different datasets. "Entity" and "Token" denote binary entity- and token-level performance. "# Sent" is unique number of sentences in training set.

Results. The evaluation results are presented in Table 7. All results are averaged over 5 runs with different random seeds. The fine-tuned RoBERTa-large model (Liu et al., 2019) achieves 86.8 and 80.2 F1 for binary tasks at token- and entity-levels. Using predictions from this model, we significantly improve existing metrics' correlation with human judgment (§4.4). We find the domain-specific models at base size, such as PubMedBERT (Tinn et al., 2021), also achieve competitive performance. However, correctly picking up the difference between seven types of complex spans remains challenging.

Transfer Learning. We use two existing datasets (Paetzold and Specia, 2016; Yimam et al., 2017) to train RoBERTa-large (Liu et al., 2019) models, and evaluated them on the testset of our MEDREADME. Table 8 presents the performance for binary complex span identification task, as existing corpora consist of binary labels, and SemEval2016 (Paetzold and Specia, 2016) only has complex word annotation. We find that both models trained using general domain data do not perform well in the medical field. This results demonstrate the necessity for our medical-focus dataset.

6 Conclusion

In this work, we present the first systematic study for sentence readability in medical domain, including a new annotated dataset, analysis, improvement to the SOTA metrics, and trained models. Our work reveal that for the specialized domain like medicine, span-level information is important for accurately rating the readability at sentence-level. By add a single feature that captures the "number of jargon" into state-of-the-art sentence readability formulas, their correlation with human judgments can be significantly improved.

Limitations

Due to the reality that major scientific medical discoveries are mainly reported in English, our

study primarily focuses on English-language medical texts. Future research could extend to medical resources in other languages. In addition, the readability ratings of a sentence can be impacted by a mixture of factors, including sentence length, grammatical complexity, word difficulty, annotator's educational background, the design and quality of annotation guidelines, as well as the target audience. We choose to use the CEFR standards, which is "the most widely used international standard" to access learners' language proficiency (Arase et al., 2022). It has detailed guidelines in 34 languages[7][8] and have been used in many prior research (Boyd et al., 2014; Rysová et al., 2016; François et al., 2016; Xia et al., 2016; Tack et al., 2017; Wilkens et al., 2018; Arase et al., 2022; Naous et al., 2023, *inter alia*). To ensure our dataset reflects the background of a broader audience, we hired U.S. college students who have completed high-school education to double annotate our MEDREADME dataset.

References

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*, 30(5):1–38.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Rosy Chhabra, Deena J Chisolm, Barbara Bayldon, Maheen Quadri, Iman Sharif, Jessica J Velazquez, Karen Encalada, Angelic Rivera, Millie Harris, Elana Levites-Agababa, et al. 2018. Evaluation of pediatric human papillomavirus vaccination provider counseling written materials: a health literacy perspective. *Academic Pediatrics*, 18(2):S28–S36.

Bernard CK Choi and Anita WP Pak. 2007. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. promotors, barriers, and strategies of enhancement. *Clinical and Investigative Medicine*, pages E224–E232.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (sle): A learned referenceless metric for sentence simplification. *arXiv preprint arXiv:2310.08170*.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021a. 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, Online. Association for Computational Linguistics.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021b. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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*

[7] http://tinyurl.com/CEFR-Standard
[8] http://tinyurl.com/CEFR-34-languages

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harika Echuri, Cole W Wendell, Symone Brown, and Mary K Mulcahey. 2022. Readability and variability among online resources for patella dislocation: What patients are reading. *Orthopedics*, 45(2):e62–e66.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

Linda Y Fu, Kathleen Zook, Zachary Spoehr-Labutta, Pamela Hu, and Jill G Joseph. 2016. Search engine ranking, quality, and content of web pages that are critical versus noncritical of human papillomavirus vaccine. *Journal of Adolescent Health*, 58(1):33–39.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Davis A Hartnett, Alexander P Philips, Alan H Daniels, and Brad D Blankenhorn. 2023. Readability and quality of online information on total ankle arthroplasty. *The Foot*, 54:101985.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Pranay Jindal and Joy C MacDermid. 2017. Assessing reading levels of health information: uses and limitations of flesch formula. *Education for Health: Change in Learning & Practice*, 30(1).

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

Erhan Kaya and Sinan Görmez. 2022. Quality and readability of online information on plantar fasciitis and calcaneal spur. *Rheumatology International*, 42(11):1965–1972.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Jessica Kruse, Paloma Toledo, Tayler B Belton, Erica J Testani, Charlesnika T Evans, William A Grobman, Emily S Miller, and Elizabeth MS Lange. 2021. Readability, content, and quality of covid-19 patient education materials from academic medical centers in the united states. *American journal of infection control*, 49(6):690–693.

Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 2022. MedJEx: A medical jargon extraction model with Wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Tarek Naous, Michael J Ryan, Mohit Chandra, and Wei Xu. 2023. Towards massively multi-domain multilingual readability assessment. *arXiv preprint arXiv:2305.14463*.

Gottfried E Noether. 1981. Why kendall tau? *Teaching Statistics*, 3(2):41–43.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Nikhil Pattisapu, Nishant Prabhu, Smriti Bhati, and Vasudeva Varma. 2020. Leveraging social media for medical text simplification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 851–860.

Michael K Rooney, Gaia Santiago, Subha Perni, David P Horowitz, Anne R McCall, Andrew J Einstein, Reshma Jagsi, and Daniel W Golden. 2021. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *Journal of patient experience*, 8:2374373521998847.

Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. 2016. Automatic evaluation of surface coherence in L2 texts in Czech. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016)*, pages 214–228, Tainan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Matthew Shardlow. 2013. The CW corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air . . . .

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, pages 4096–4102.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Jeniya Tabassum, Wei Xu, and Alan Ritter. 2020. WNUT-2020 task 1 overview: Extracting entities and relations from wet lab protocols. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 260–267, Online. Association for Computational Linguistics.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rodrigo Wilkens, Leonardo Zilio, and Cédrick Fairon. 2018. SW4ALL: a CEFR classified and aligned corpus for language learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Meg Wilson. 2009. Readability and patient education materials used for low-income populations. *Clinical Nurse Specialist*, 23(1):33–40.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Zhan Xu, Lauren Ellis, and Laura R Umphrey. 2019. The easier the better? comparing the readability and engagement of online pro-and anti-vaccination articles. *Health education & behavior*, 46(5):790–797.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 - complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *Biological and Medical Data Analysis: 6th International Symposium, ISBMDA 2005, Aveiro, Portugal, November 10-11, 2005. Proceedings 6*, pages 184–192. Springer.

A Formulas of Readability Metrics

In this section, we list the formulas for four unsupervised readability metrics.

FKGL. The Flesch-Kincaid Grade Level formula is a well-known readability test designed to indicate how difficult a text in English is to understand. It is calculated using the formula:

$$FKGL = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

ARI. The Automated Readability Index (ARI) is another widely used readability metric that estimates the understandability of English text. It is formulated based on characters rather than syllables. The ARI formula is given by:

$$ARI = 4.71\left(\frac{\text{total characters}}{\text{total words}}\right) + 0.5\left(\frac{\text{total words}}{\text{total sentences}}\right) - 21.43$$

SMOG. The SMOG (Simple Measure of Gobbledygook) Index is a readability formula that measures the years of education needed to understand a piece of writing. SMOG is particularly useful for higher-level texts. The formula is as follows, where the polysyllables are calculated by counting the number of words in a text that have three or more syllables:

$$P = \text{number of polysyllables}$$
$$S = \text{number of sentences}$$
$$SMOG = 1.0430\sqrt{P \times \frac{30}{S}} + 3.1291$$

RSRS. The RSRS (Ranked Sentence Readability Score) leverages log probabilities from a neural language model and the sentence length feature. It's calculated through a weighted sum of individual word losses. Each word's Negative Log Loss (WNLL) is sorted in ascending order and weighted by its rank. The formula assigns higher weights to the out-of-vocabulary (OOV) words, by setting $\alpha = 2$ for all OOV words and 1 for others. The formula for RSRS is:

$$RSRS = \frac{\sum_{i=1}^{S}[\sqrt{i}]^{\alpha} \cdot WNLL(i)}{S}$$

And WNLL can be calculated by:

$$WNLL = -(y_t \log y_p + (1 - y_t)\log(1 - y_p))$$

Here, y_p is the predicted distribution from the language model, and y_t is the empirical distribution, where 1 for words that appear in the text, and 0 for all others.

B Introduction of Medical Text Simplification Resources

Our dataset is constructed on top of open-accessed resources. Each of the resources is detailed below.

Biomedical Journals. The latest advancements in the medical field are documented in the research papers. To improve accessibility, the authors or domain experts sometimes write a summary in lay language, providing a valuable resource for studying medical text simplification. We include five sub-journals from NIHR, five sub-journals from PLOS, and the Proceedings of the National Academy of Sciences (PNAS) compiled by (Guo et al., 2022). In addition, we also include the eLife corpus compiled by (Goldsack et al., 2022), which consists of the paper abstracts and summaries in life sciences written by expert editors.

Cochrane Reviews. As "the highest standard in evidence-based healthcare", Cochrane Review[9] provides systematic reviews for the effectiveness of interventions and the quality of diagnostic tests in healthcare and health policy areas, by identifying, appraising, and synthesizing all the empirical evidence that meets pre-specified eligibility criteria. We use the parallel corpus compiled by (Devaraj et al., 2021a).

Medical Wikipedia. As their original and simplified versions are created independently in a collaboration process, the two versions are on the same topic but may not be entirely aligned (Xu et al., 2015). We apply the state-of-the-art methods (Jiang et al., 2020) to extract aligned paragraph pairs from Wikipedia, of which we improve the quality and quantity over existing work (Pattisapu et al., 2020). Specifically, we first collect 60,838 medical terms using Wikidata's SPARQL service[10] by querying unique terms that have 30 specific properties, including UMLS code, medical encyclopedia, and the ontologies for disease, symptoms, examination,

[9] https://www.cochranelibrary.com/
[10] https://query.wikidata.org/

drug, and therapy. Then, we extract corresponding articles for each term from Wikipedia and simple Wikipedia dumps,[11] based on title matching using WikiExtractor library,[12] resulting in 2,823 aligned article pairs after filtering the empty pages. Finally, we use the state-of-the-art neural CRF sentence alignment model (Jiang et al., 2020) with 89.4 F1 on Wikipedia to perform paragraph and sentence alignment for each complex-simple article pair.

Merck Manuals. We use the segment pairs from prior work (Cao et al., 2020), which are manually aligned by medical experts.

C Implementation Details for Complex Span Identification Models

We use the Huggingface[13] implementations of the BERT and RoBERTa models. We tune the learning rate in {1e-6, 2e-6, 5e-6, 1e-5, 2e-5} based on F1 on the devset, and find 2e-6 works best for our best performing RoBERTa-large model. All models are trained within 1.5 hours on one NVIDIA A40 GPU.

D More Results for Complex Span Identification

Table 9 presents the results of the exact match at entity level for the complex span identification task on the MEDREADME testset. As medical jargon and complex spans have diverse formats in the medical articles, it is challenging for the models to predict the exact matched entities.

Models	Binary	3-Class	7-Category
Large-size Models			
BERT (2019)	72.0	68.2	48.5
RoBERTa (2019)	**74.9**	**71.2**	**64.1**
BioBERT (2020)	72.4	67.6	60.5
PubMedBERT (2021)	73.4	69.9	62.2
Base-size Models			
BERT (2019)	70.7	67.0	59.3
RoBERTa (2019)	73.5	70.0	62.4
BioBERT (2020)	70.5	67.1	59.8
PubMedBERT (2021)	72.2	69.0	61.2

Table 9: **Micro F1** of exact match at entity-level for complex span identification task on the MEDREADME testset. The **best** and second best scores within each model size are highlighted. Models are trained with fine-grained labels in seven categories and evaluated at different granularity.

[11] The March 22, 2023 version.
[12] https://attardi.github.io/wikiextractor/
[13] https://github.com/huggingface/transformers

E Prompts for Sentence Readability

Rate the following sentence on its readability level. The readability is defined as the cognitive load required to understand the meaning of the sentence. Rate the readability on a scale from very easy to very hard. Base your scores on the CEFR scale for L2 learners. You should use the following key:

1 = Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.

2 = Can understand short, simple texts on familiar matters of a concrete type

3 = Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.

4 = Can read with a large degree of independence, adapting style and speed of reading to different texts and purpose

5 = Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.

6 = Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings.

EXAMPLES:

Sentence: "[EXAMPLE 1]"

Given the above key, the readability of the sentence is (scale=1-6): [RATING 1]

Sentence: "[EXAMPLE 2]"

Given the above key, the readability of the sentence is (scale=1-6): [RATING 2]

Sentence: "[EXAMPLE 3]"

Given the above key, the readability of the sentence is (scale=1-6): [RATING 3]

Sentence: "[EXAMPLE 4]"

Given the above key, the readability of the sentence is (scale=1-6): [RATING 4]

Sentence: "[EXAMPLE 5]"

Given the above key, the readability of the sentence is (scale=1-6): [RATING 5]

Sentence: "[TARGET SENTENCE]"

Given the above key, the readability of the sentence is (scale=1-6): [RATING]

Table 10: Following (Naous et al., 2023) in prompt construction, we utilize the same description of the six CEFR levels that were provided to human annotators, along with five examples and their ratings, randomly sampled from the dev set. Then, the model is instructed to evaluate the readability of a given sentence. The full template is presented above.

F Annotation Interface for Sentence Readability

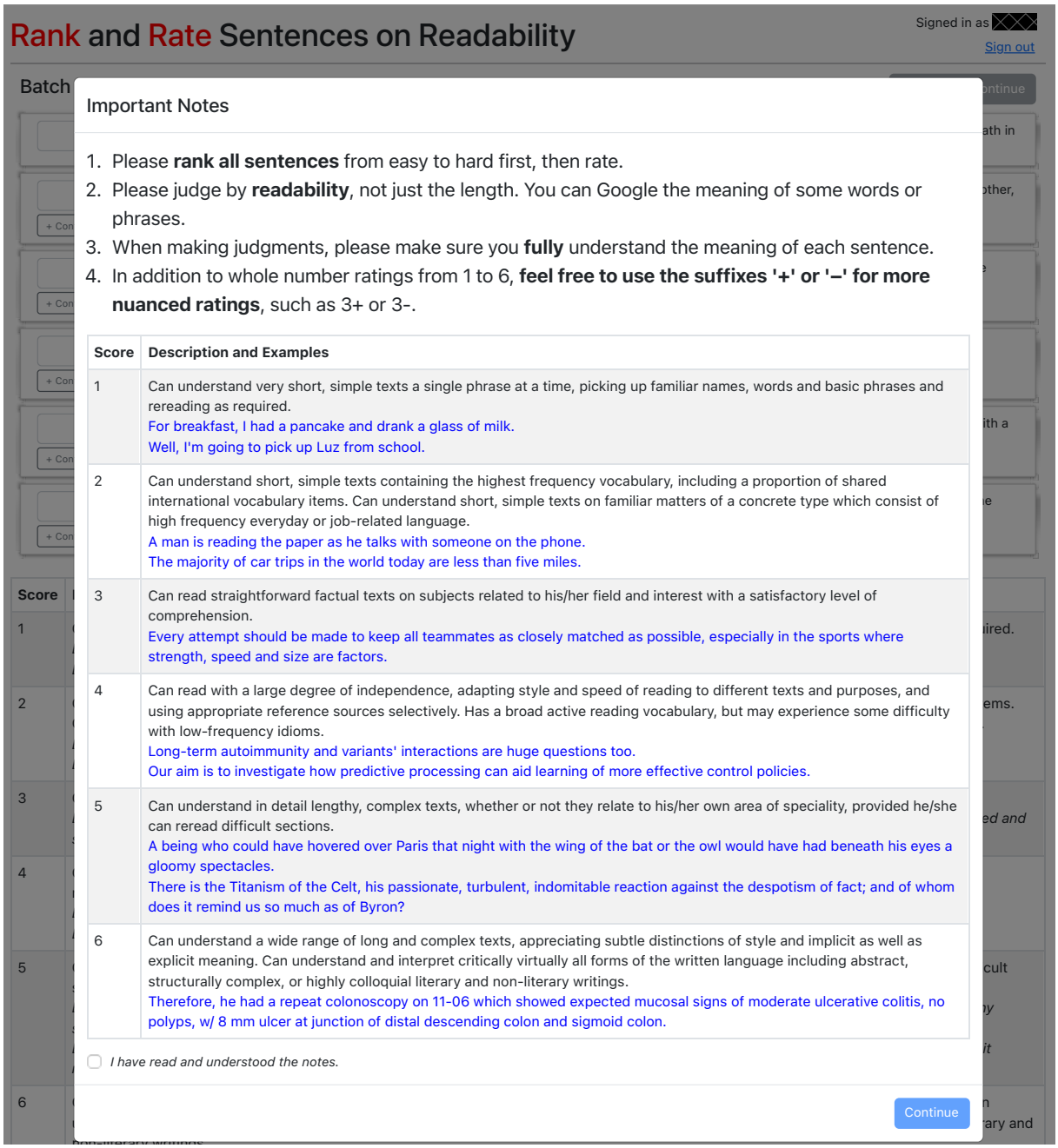
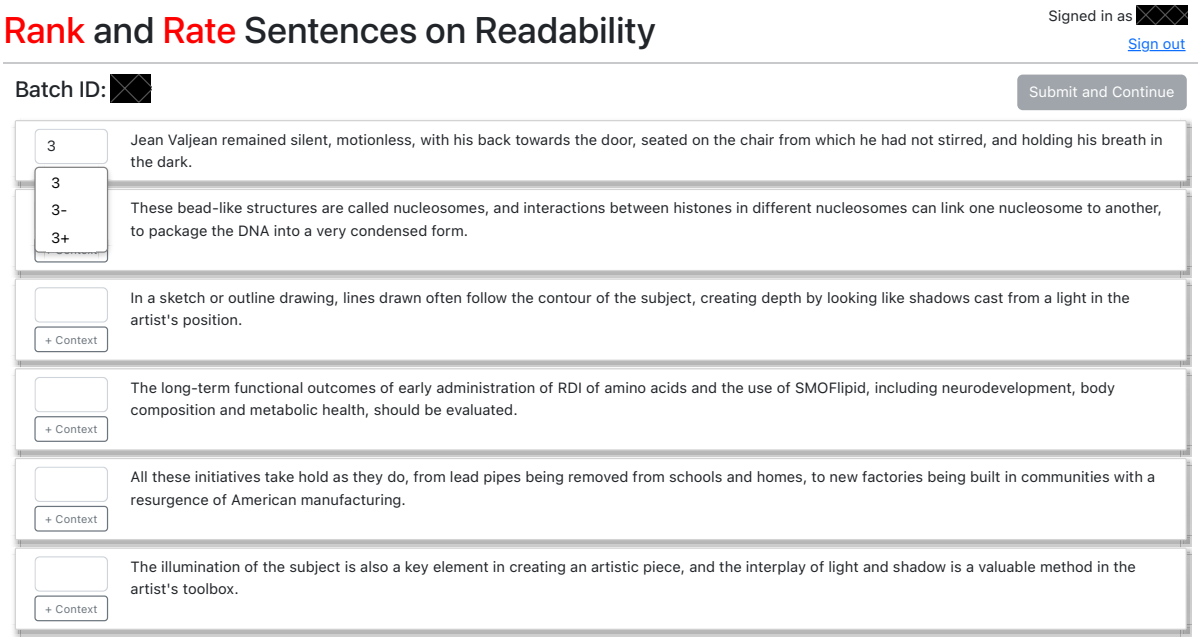


Figure 5: Instructions for annotating the sentence readability.



Score	Description and Examples
1	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required. *Example: For breakfast, I had a pancake and drank a glass of milk.* *Example: Well, I'm going to pick up Luz from school.*
2	Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items. Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language. *Example: A man is reading the paper as he talks with someone on the phone.* *Example: The majority of car trips in the world today are less than five miles.*
3	Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension. *Example: Every attempt should be made to keep all teammates as closely matched as possible, especially in the sports where strength, speed and size are factors.*
4	Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms. *Example: Long-term autoimmunity and variants' interactions are huge questions too.* *Example: Our aim is to investigate how predictive processing can aid learning of more effective control policies.*
5	Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections. *Example: A being who could have hovered over Paris that night with the wing of the bat or the owl would have had beneath his eyes a gloomy spectacles.* *Example: There is the Titanism of the Celt, his passionate, turbulent, indomitable reaction against the despotism of fact; and of whom does it remind us so much as of Byron?*
6	Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning. Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. *Example: Therefore, he had a repeat colonoscopy on 11-06 which showed expected mucosal signs of moderate ulcerative colitis, no polyps, w/ 8 mm ulcer at junction of distal descending colon and sigmoid colon.*

Figure 6: The interface for annotating sentence readability. Annotators can click the "+ Context" button to see the surrounding sentences.

G Annotation Interface for Complex Span Identification

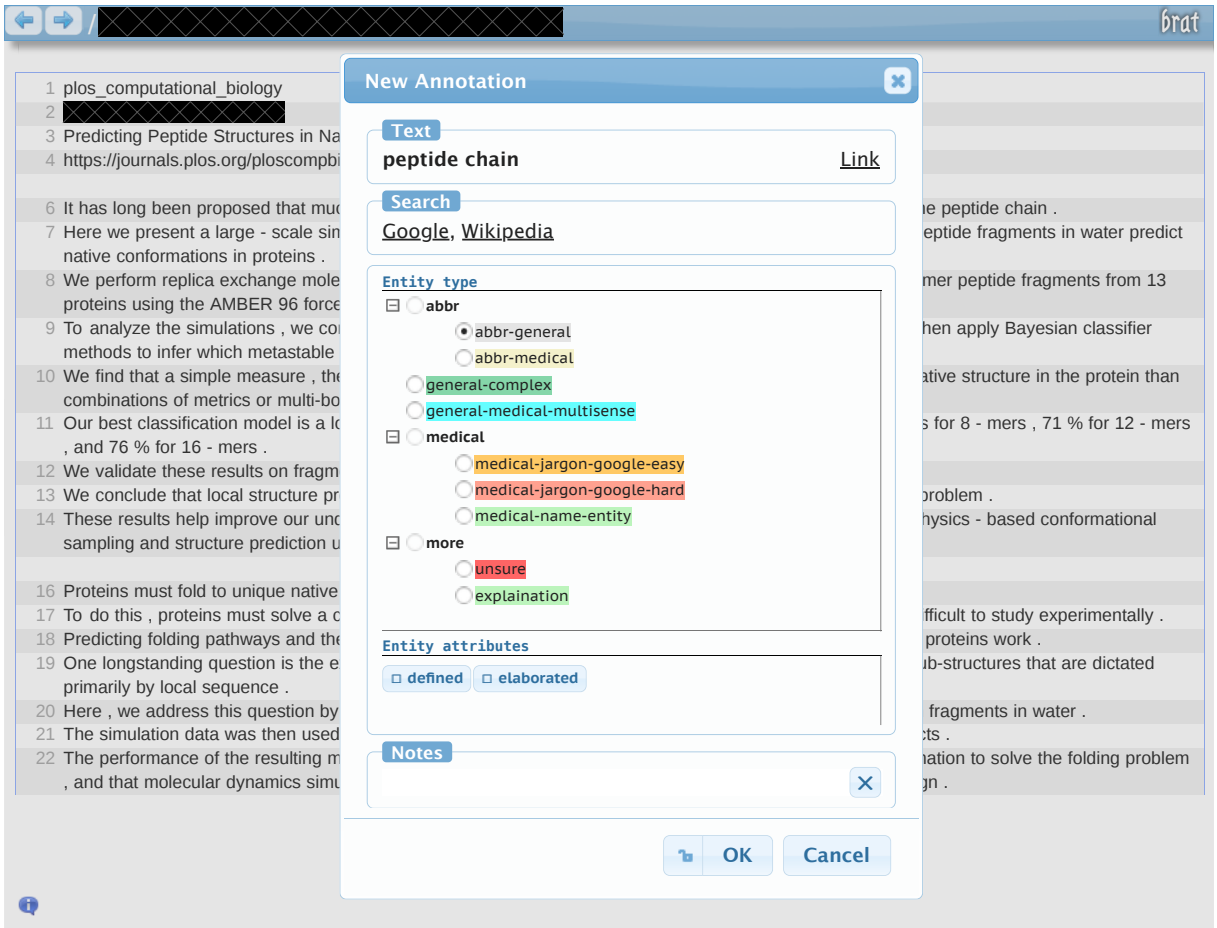


Figure 7: The annotation interface for complex span identification.