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

Chao Jiang
College of Computing
Georgia Institute of Technology
chaojiang@gatech.edu

Wei Xu
College of Computing
Georgia Institute of Technology
wei.xu@cc.gatech.edu

Abstract

Medical texts are notoriously challenging to read. Properly measuring their readability is the first step towards making them more accessible. In this paper, we present a systematic study on fine-grained readability measurements in the medical domain at both sentence-level and span-level. We introduce a new dataset MedReadMe, which consists of manually annotated readability ratings and fine-grained complex span annotation for 4,520 sentences, featuring two novel "Google-Easy" and "Google-Hard" categories. It supports our quantitative analysis, which covers 650 linguistic features and automatic complex word and jargon identification. Enabled by our high-quality annotation, we benchmark and improve several state-of-the-art sentence-level readability metrics for the medical domain specifically, which include 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 readability formulas can significantly improve their correlation with human judgments. We will publicly release the 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 Wikipedia contain extensive medical information, while research papers introduce the latest findings, including emerging medical conditions and treatments (Joseph et al., 2023). However, comprehending these resources can be very challenging due to their technical nature and the extensive use of specialized terminology (Zeng et al., 2005). As the first step to making

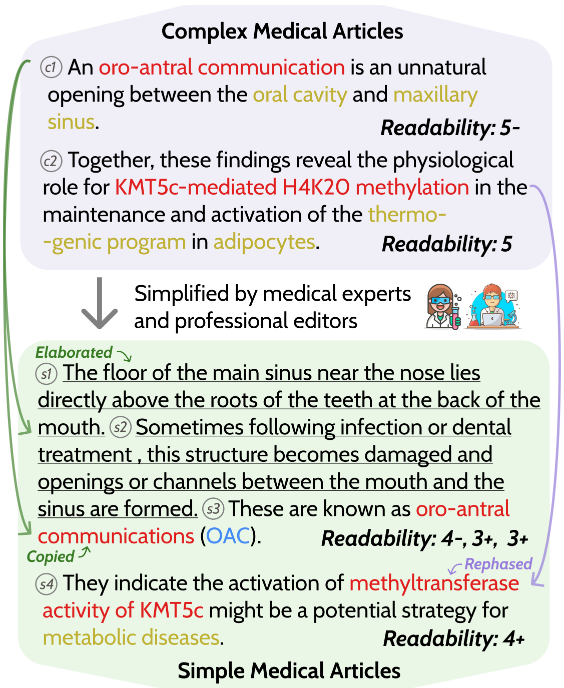


Figure 1: An illustration of our dataset, with sentence readability ratings and fine-grained complex span annotation on 4,520 sentences, including "Google-Hard" and "Google-Easy", abbreviations, and general complex terms, etc. We also analyze how medical jargon are being handled during simplification.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

them more accessible, properly measuring the readability of medical texts is crucial (Rooney et al., 2021; Echuri et al., 2022). However, a high-quality multi-source dataset for reliably evaluating and improving sentence readability metrics for medical domain is lacking.

To address this gap in research, we present a systematic study for medical text readability in this paper, which includes a manually annotated readability dataset (§2), a data-driven analysis to answer "*why medical sentences are so hard*", covering 650 linguistic features and additional medical jargon features (§3), a comprehensive benchmark of state-of-the-art readability metrics (§4.1), a simple yet

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17293–17319
November 12-16, 2024 ©2024 Association for Computational Linguistics

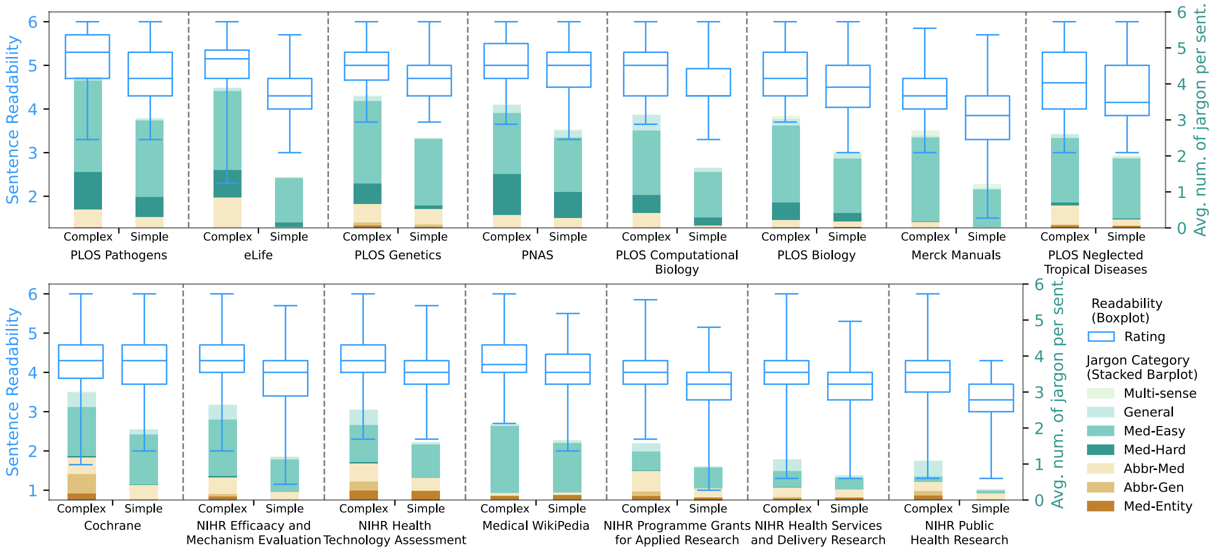


Figure 2: The distribution of sentence readability (boxplot on the left y-axis) and the average number of jargon spans per category (stacked barplot on the right y-axis) in each sentence across both "complex" and "simplied" versions for 15 commonly used resources for medical text simplification. Sentences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

effective method to improve LM-based readability metrics by training on our dataset (§4.2), and an automatic model that can identify complex words and jargon with fine-grained categories (§5).

Our MEDREADME dataset consists of 4,520 sentences with both sentence-level readability ratings and fine-grained complex span-level annotations (Figure 1). It covers complex-simple parallel article pairs from 15 diverse data resources that range from encyclopedias to plain-language summaries to biomedical research publications (Figure 2). 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 other methods (Naous et al., 2023). We also ask lay annotators to highlight any words/phrases that they find hard to understand and categorize the reason using a 7-class taxonomy. Considering that "the majority of people seek health information online began at a search engine",[1] we introduce two categories of "Google-Easy" and "Google-Hard" to reflect whether jargon is understandable after a quick Google search, providing a fresh perspective beyond binary or 5-point Likert scales.

Our new dataset addresses three limitations in prior work: (1) Existing work with sentence-level ratings mainly covers data from general domains, such as Wikipedia (De Clercq and Hoste, 2016), news (Stajner 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). (2) Prior work separates the research on sentence readability and complex jargon terms, hence missing the possible correlations between them (Kwon et al., 2022; Naous et al., 2023). (3) Previous research on sentence readability uses document-level ratings as an approximation, which is shown to be inaccurate (Arase et al., 2022; Cripwell et al., 2023).

Our analysis reveals that compared to various linguistic features, complex spans, especially medical jargon from certain domains, more significantly elevate the difficulty of sentences (§3.1). We also scrutinize the quality of 15 widely used medical text simplification resources (§3.4), and find that there are non-negligible variances in readability among them, as shown by the differences in the height of the box plots in Figure 2. While evaluating various sentence readability metrics, we find that unsupervised methods based on lexical features perform poorly in the medical domain. Prompting large language models such as GPT-4 (Achiam et al., 2023) with 5-shot achieves strong performance, yet is outperformed by fine-tuned models in a much smaller size. Inspired by our analysis, 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 performance and also make them more stable.

2 Constructing MEDREADME Corpus

This section presents the detailed procedure for constructing the Medical Readability Measurement (MEDREADME) corpus, which consists of 4,520 sentences in 180 complex-simple article pairs randomly sampled from 15 data sources (§2.1).

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

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 1: A taxonomy ($\mathcal{I}$) of complex textual spans in the medical domain with examples highlighted by a red background. The “Medical Jargon” and "Abbreviation" rows are based on the aggregation of sub-categories.

2.1 Data Collection and Preprocessing

Different from prior work (Arase et al., 2022; Naous et al., 2023), our study consists of sentences from complete complex-simple article pairs, enabling a deeper analysis of how professional editors simplify medical documents. The 15 resources that we considered include (1) the abstract sections and plain-language summaries from scientific papers, such as the National Institute for Health and Care Research (NIHR) and Cochrane Review of “the highest standard in evidence-based healthcare”,[2] for which we use the aligned article pairs released from prior studies (Devaraj et al., 2021a; Goldsack et al., 2022; Guo et al., 2022); and (2) segment and paragraph pairs in the parallel versions of medical references from trusted online platforms, such as Merck Manuals[3] and medical-related Wikipedia articles we extracted. A detailed description of each resource and pre-processing steps is provided in Appendix C.

Target Audience. To ensure our study reflects the background of a broader audience, our study mainly targets people who have completed high school or are entering college, and our dataset is annotated by college students without medical backgrounds using a six-point Likert scale.

2.2 Sentence-level Readability Annotation

To collect ground-truth judgments, we hire three university students with prior linguistic annotation experience to annotate the readability ratings for 4,520 sentences. We utilize the “rank-and-rate” interface (Naous et al., 2023) and the CEFR scale (Arase et al., 2022), with several improvements.

Annotation Guidelines. Following prior work (Arase et al., 2022), we adopt the Common European Framework of Reference for Languages (CEFR) to annotate the sentence readability. CEFR standards were originally created for language learners. Because the scale is essentially a six-point Likert scale, we believe the findings would be mostly generalizable to a broader audience, including native speakers. Another reason for using the CEFR scale is to make our work comparable to the existing work and datasets which were created using the CEFR standards.

CEFR Scale. CEFR is the most widely used international criteria to define learners’ language proficiency, assessing language skills on a 6-level scale with detailed guidelines,[4] from beginners (A1) to advanced mastery (C2), which are denoted as level 1 (easiest) to level 6 (hardest) in our interface. Following prior work (Arase et al., 2022; Naous et al., 2023), a sentence’s readability is determined based on the CEFR level, at which an individual can understand the sentence without assistance. As medical texts naturally concentrate on the harder-to-understand side, we introduce the use of “+” and “-” signs to differentiate the nuance in readability, e.g., “3+” and “3-”, in addition to each integer level. They are treated as 3.3 and 2.7 when converting to the numeric scores.

[2] https://www.cochranelibrary.com/
[3] https://www.merckmanuals.com/
[4] https://tinyurl.com/CEFR-Standard/

Feature	Corr.
Number of unique sophisticated lexical words†	0.645
Corrected type-token-ratio (CTTR)	0.627
Number of syllables	0.589
Max age-of-acquisition (AoA) of words (2012)	0.576
Number of unique words	0.574
Number of words	0.532
Average number of characters per token	0.524
Corrected noun variation	0.513
The maximum dependency tree depth	0.437
Cumulative Zipf score for all words (2012)	0.425

Table 2: Top representative linguistic features and their Pearson correlation with readability. †Sophisticated lexical words (Lu, 2012) are nouns, non-auxiliary verbs, adjectives, and certain adverbs that are not in the 2,000 most frequent lemmatized tokens in the `American National Corpus` (ANC). More features and more implementation details are provided in the Appendix B.

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

Quality Control. For each medical sentence we annotate for the MedReadMe corpus, we sample another (mostly non-medical) sentence with comparable length from the existing ReadMe++ dataset (Naous et al., 2023) as a "control". Therefore, each set of sentences shown to the annotator consists of three medical sentences and three control sentences whose ratings are known. Annotators are asked to spend at least three minutes on every set, and their annotation quality is monitored through the use of control sentences. The 1,924 sentences in the dev and test sets are double annotated, and the scores are merged by average. The inter-annotator agreement is 0.742 measured by Krippendorff's alpha (Krippendorff, 2011). On the control sentences, our annotation achieves a Pearson correlation of 0.771 with the original ratings from ReadMe++.

2.3 Fine-trained Complex Span Annotation

We propose a new taxonomy to comprehensively capture 7 different categories of complex spans that appeared in the medical texts, as shown in Table 1. The complete annotation guideline with more examples is provided in Appendix L.

Type	#Spans	#Tokens	%Tokens
Medical Jargon	0.644	0.591	0.445
Abbreviation	0.259	0.254	0.134
General Complex	0.112	0.09	0.001
Multi-sense	0.058	0.059	0.035
All Categories	0.656	0.617	0.584

Table 3: The impact of 15 features related to complex spans, measured by the Pearson correlation with ground-truth sentence readability on the MedReadMe dataset.

"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 seemingly minor distinction can have great implications in developing technological advances for medical text simplification and health literacy, motivating us to propose a novel category "Google-Hard" for medical jargon, which is separate from jargon that is "Google-Easy" or "Name-Entity". In total, our dataset captures 698 Google-Hard medical jargon and 5,251 Google-Easy ones.

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

3 Key Findings

Enabled by our MedReadMe corpus, we first analyze the sentence readability measurements for medical texts (§3.1 and §3.4), then dive into medical jargon of different complexities (§3.2 and §3.3).

3.1 Why Medical Texts are Hard-to-Read?

The readability of a sentence can be impacted by a mixture of factors, including sentence length, grammatical complexity, word choice, etc. We extract 650 linguistic features from each sentence and measure their correlation with ground-truth readability.

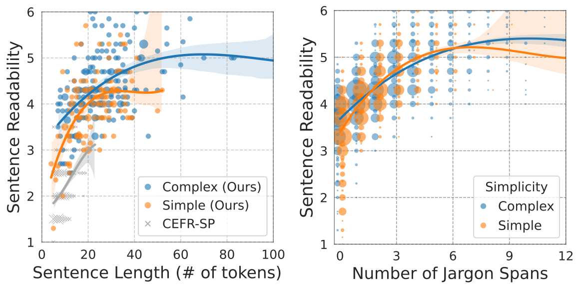


Figure 3: *Left*: Readability of sentences with different lengths. Compared to the CEFR-SP dataset (Arase et al., 2022), our corpus contains much longer sentences. *Right*: Readability of sentences with different numbers of jargon. The circle's radius reflects the number of overlapping points at each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

15 additional features are designed to quantify the influence of complex spans. Based on our qualitative analysis, we found that complex spans, such as medical jargon, have a more profound impact on readability compared to other linguistic aspects.

Impact of linguistic features. For each sentence, 650 linguistic features are extracted, including syntax and semantics features, quantitative and corpus linguistics features, in addition to psycholinguistic features (Vajjala and Meurers, 2016), such as the age of acquisition (AoA) released by Kuperman et al. (2012), and concreteness, meaningfulness, and imageability extracted from the MRC psycholinguistic database (Wilson, 1988). These features are extracted using a combination of toolkits, each of which covers a different subset of features, including `LFTK` (Lee and Lee, 2023), `LingFeat`, `Profiling-UD` (Brunato et al., 2020a), `Lexical Complexity Analyzer` (Lu, 2012), and `L2 Syntactic Complexity Analyzer` (Lu, 2010). We select and present top-10 representative features in Table 2, and provide a more complete list of the top-50 influential features in Appendix B with more detailed definition of each feature. We found that resource-based methods, such as the count of "sophisticated lexical words" (Lu, 2012) and Zipf score (Powers, 1998), are very useful. Length-related features are also informative.

Impact of Complex Spans. Based on our pilot study and feedback from annotators, we observed that the specialized terminology, while allowing for precise and concise communication among experts, significantly affects the difficulty level of texts in specialized domains. With our fine-grained span-level annotations (§2.3), we can directly measure the effects that each type of

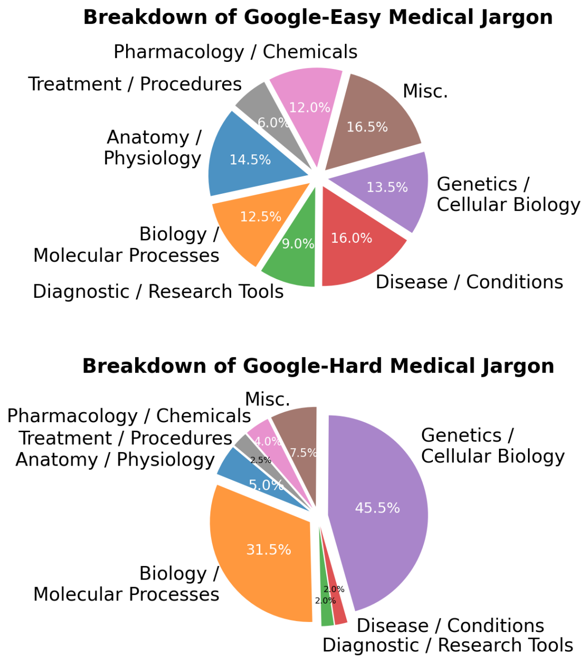


Figure 4: Breakdown of Google-Easy and Google-Hard jargon into different medical domains based on our manual analysis of 400 randomly sampled jargon.

complex words and jargon have on readability. Specifically, we design three features "number-of-jargon-spans", "number-of-jargon-tokens", and "percentage-of-jargon-tokens" for complex span in each category: *medical jargon*, *abbreviation*, *general complex terms*, and *multi-sense words*. We then compute their correlation with the sentence-level readability ratings. As shown in Table 3, we find that medical jargon significantly affects readability, and abbreviations follow in influence.

Figure 3 plots the relationship between readability and both sentence length (*left*) and the number of jargon spans (*right*). On the left, we notice that the lines representing "complex" and "simple" sentences begin to diverge as sentence length exceeds 20 tokens, suggesting that factors beyond length affect the readability. In contrast, a stronger overall correlation between the number of jargon spans and readability is observed in the right figure.

3.2 What Makes a Jargon Easy (or Hard)?

Based on the feedback from annotators, we identify two major factors that influence the perceived difficulty of medical jargon, as listed below:

Inherent Complexity of Topics. To analyze the perceived difficulty of medical jargon from different domains, we randomly sample 200 Google-Easy and 200 Google-Hard medical jargon, and manually analyze their topics. The results are presented in Figure 4. Google-Easy terms are more

Sources	Length	FKGL (Kincaid et al.)	ARI (Smith and Senter)	SMOG (Mc Laughlin)	RSRS (Martinc et al.)	FKGL-Jar (Ours)	ARI-Jar (Ours)	SMOG-Jar (Ours)	RSRS-Jar (Ours)
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 4: 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 adding a "number-of-jargon" feature into existing readability formula (more details in §4.2). Our proposed method significantly improves the correlation over existing metrics, as demonstrated by the average correlation.

Operation	Google-Easy	Google-Hard
Knowledge Panel		
Covered	45.6%	10.3%
Explained by Figure	13.6%	4.6%
Feature Snippets		
Covered	55.3%	21.2%
Highlighted Text	52.4%	18.5%
Explained by Figure	22.8%	3.6%

Table 5: The percentage of explanatory content provided by Google. An annotated screenshot of the webpage is provided in Figure 6 in Appendix I to visually demonstrates *"Knowledge Panel"* and *"Feature Snippets"*,

diversified across different topics, while Google-Hard terms mainly fall under *Genetics / Cellular Biology* and *Biology / Molecular Processes*. This suggests that jargon associated with genetics or molecular procedures tends to be more challenging to read, possibly due to the specialized knowledge required to interpret them.

Variance in the Explanation. We also observed that the accessibility of medical jargon is greatly improved when search engines offer explanations or visual aids in their results. Search engines may provide the explanation of a medical term in two places: (1) the feature snippets in the answer box; and (2) the knowledge panel, which is powered by a knowledge graph. An annotated screenshot of the search results is provided in Figure 6 in Appendix I to demonstrate each element visually. By parsing the Google search results for 2,731 unique Google-Easy and 504 Google-Hard medical jargon from our corpus, we quantified the existence of these explanations in Table 5. The Google-Easy jargon is more frequently accompanied by explanatory content compared to the Google-Hard category. The use of visual aids also follows a similar pattern; Google-Easy terms are much more likely to be explained by figures compared to Google-Hard ones.

Operation	Google-Easy	Google-Hard
Kept	22%	13% (↓ 9%)
Deleted	56%	52% (↓ 4%)
Rephrased	3%	10% (↑ 7%)
Kept + Explained	8%	8% (–)
Del.+ Explained	11%	17% (↑ 6%)

Table 6: The distribution of operations to 200 medical jargon (100 in each type), based on our manual analysis.

3.3 How Professional Editors Simplify the Medical Jargon?

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 6. 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 with different complexities.

3.4 Readability Significantly Varies Across Existing Medical Simplification Corpora

To better understand the quality of medical text simplification corpora, in Figure 2, we plot the distribution of sentence readability and numbers of jargon per sentence across 15 different resources. Within each source, the simplified texts are rated as easier to understand than their complex counterparts, though the extent varies. However, when compared across the board, simplified texts from some sources can be even more challenging to read than the complex texts from other sources, suggesting that not all plain texts are equally simple. In addition, some resources, such as "PLOS pathogens", are especially difficult for laypersons without domain-specific knowledge to understand. The current research practice in medical text simplification often treat all data uniformly, such as concatenating all available corpora into one 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 readability level and their use cases.

4 Medical Readability Prediction

In this section, we present a comprehensive evaluation of state-of-the-art readability metrics for medical texts (§4.1), and design a simple yet effective method to further improve them (§4.2).

4.1 Evaluating Existing Readability Metrics

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

Unsupervised Methods. The Pearson correlations between ground-truth readability and each unsupervised metric are presented in the left half of Table 4. The metrics we considered include FKGL (Kincaid et al., 1975), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), and RSRS (Martinc et al., 2021), and their detailed formulations are provided in Appendix A. We also add sentence length as a baseline. We find that the unsupervised methods generally do not perform very well. The language model-based RSRS score significantly outperforms the traditional feature-based metrics, among which SMOG performs best.

Supervised and Prompt-based Methods. The results are presented in Table 7. For supervised methods, we fine-tune language models on our dataset and 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 GPT-4 and Llama-3[5] (AI@Meta, 2024) using 5-shot. The prompts are constructed following Naous et al. (2023). More details and the full prompt template are in Appendix H. We find that prompt-based methods achieve competitive results, e.g., GPT-4 outperforms the strongest unsupervised metric RSRS, although they still fall behind supervised methods.

[5]More specifically, we used `gpt-4-0613` and `Llama-3.1-8B-Instruct` in the experiments.

4.2 Improving Readability Metrics with Jargon Identification

To incorporate the consideration of jargon into existing metrics, we add and tune a weight α for the feature "number-of-jargon" as follows:

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

where "FKGL-Jar" denotes adding jargon into the FKGL score, similarly for other metrics with a suffix "-Jar". The weight α is chosen by grid search on the dev set using gold annotation for each metric. As RSRS scores are smaller than 1, we scale them by 100 before the parameter search. The right sides in Table 4 and 7 report the performance of each unsupervised and supervised method on the test set, after adding our proposed term. To reflect the real-world 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 (α) 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.

Length-Controlled Experiment. To analyze the impact on sentences of varied lengths, in Figure 5, we present the 95% confidence intervals for the Kendall Tau-like correlation (Noether, 1981) between the ground-truth readability and predictions from each metric (Maddela et al., 2023). We find the proposed "-Jar" term is advantageous for sentences at all lengths and is especially helpful for feature-based methods, such as SMOG. In addition, the incorporation of jargon makes the metrics more stable, as demonstrated by the narrower intervals.

5 Fine-grained Complex Span Identification

Based on our analysis in §4.2, 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, 2019), and utilize our annotated dataset to train and evaluate several models.

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 dif-

Sources	5-shots		Trained on Each Corpus				The Trained + an Jargon Term			
	GPT-4 (Achiam et al.)	Llama 3-8b (AI@Meta)	ReadMe++ (Naous et al.)	CEFR-SP (Arase et al.)	CompDS (Brunato et al.)	MEDREADME (Ours)	$\text{ReadMe++}_{\text{Jar}}$ (Ours)	$\text{CEFR-SP}_{\text{Jar}}$ (Ours)	$\text{CompDS}_{\text{Jar}}$ (Ours)	$\text{MEDREADME}_{\text{Jar}}$ (Ours)
Cochrane	0.908	0.665	0.858	0.899	0.870	0.947	0.842	0.850	0.785	0.882
PNAS	0.780	0.528	0.852	0.820	0.791	0.874	0.780	0.824	0.744	0.873
NIHR Series	0.713	0.485	0.824	0.753	0.706	0.885	0.697	0.687	0.634	0.700
eLife	0.538	0.188	0.594	0.715	0.608	0.712	0.812	0.802	0.777	0.861
PLOS Series	0.672	0.520	0.680	0.691	0.635	0.702	0.787	0.843	0.744	0.850
Wiki	0.670	0.447	0.824	0.709	0.607	0.843	0.712	0.619	0.673	0.709
MSD	0.766	0.562	0.784	0.778	0.757	0.867	0.918	0.880	0.863	0.937
Mean ± Std	0.721 ± 0.115	0.485 ± 0.148	0.774 ± 0.1	0.766 ± 0.073	0.711 ± 0.101	0.833 ± 0.092	0.793 ± 0.076	0.786 ± 0.096	0.746 ± 0.075	0.830 ± 0.090

Table 7: Pearson correlation (↑) between human ground-truth readability and each **prompting** and **supervised** readability metric. 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 §4.2). Prompt-based methods are competitive, while still outperformed by fine-tuned models in much smaller sizes.

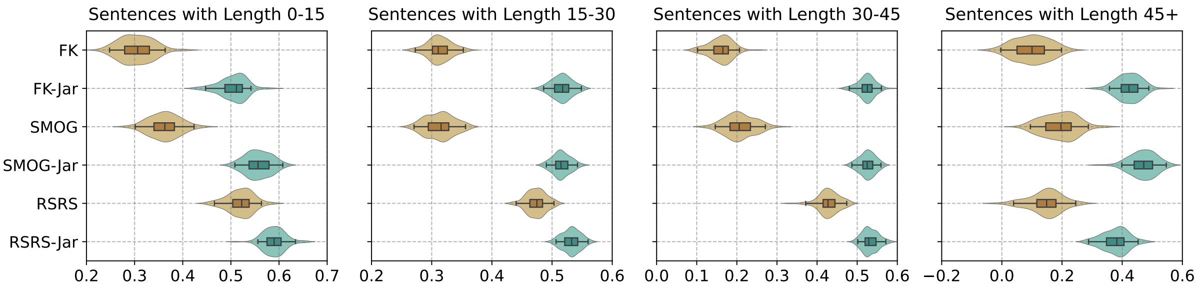


Figure 5: The 95% confidence intervals for Kendall Tau-like correlation (↑) between ground-truth readability annotation and predicted outputs from each automatic metric for sentences with different lengths, calculated by bootstrapping (Deutsch et al., 2021). In addition to a higher correlation with human judgments, incorporating jargon ("-Jar") makes each metric more stable, as shown by the smaller intervals.

ferent pre-trained embeddings. The implementation details are provided in Appendix D.

Evaluation Metrics. We consider two variants of F1 measurements: (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 span. We use the evaluation script released by Tabassum et al. (2020).[6] We also report the exact match performance at entity-level in the Appendix F. (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 7 categories, (2) associated 3 higher-level classes (i.e., medical / general+multisense / abbreviation), and (3) binary judgments between complex or non-complex text spans.

Results. The evaluation results are presented in Table 8. 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 readability metrics' correlation with human judgment (§4.2). We find the domain-specific models at base size, such as PubMedBERT (Tinn et al., 2021), also achieve competitive performance. However, differentiating between the seven categories of complex spans remains challenging.

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**	<u>68.3</u>	79.0	<u>75.2</u>	66.5
Base-size Models						
BERT (2019)	85.4	80.4	66.3	77.0	72.5	63.3
RoBERTa (2019)	<u>86.2</u>	<u>81.7</u>	68.0	<u>79.7</u>	75.2	<u>66.6</u>
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 8: **Micro F1** (↑) of different systems for complex span identification on the MEDREADME test set. The **best** and <u>second-best</u> scores are highlighted. Models are trained with fine-grained labels in seven categories and evaluated at different granularity.

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 test set of our MEDREADME. Table 9 presents the performance for binary com-

[6] https://github.com/jeniyat/WNUT_2020_NER/tree/master/code/eval

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

plex 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 Related Work

Readability Measurement in Medical Domain. Unsupervised metrics, such as FKGL (Kincaid et al., 1975), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), and Coleman-Liau index (Coleman and Liau, 1975) have been widely adopted in existing research on the medical readability analysis, as they do not require training data (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 shallow lexical features. Unsupervised 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), while other supervised metrics rely on fine-tuning LLMs on the annotated corpora (Arase et al., 2022; Naous et al., 2023); however, previously, the performance of these methods on the medical texts were unclear. Enabled by our high-quality dataset, we benchmark existing state-of-the-art metrics in the medical domain (§4.1), and also further improve their performances (§4.2).

Complex Span Identification in Medical Domain. Kauchak and Leroy (2016) collects a dataset that consists of the difficulty for 275 words. CompLex 2.0 (Shardlow et al., 2020) consists of complex spans rated on a 5-point Likert scale. However, it only covers spans with one or two tokens. MedJEx corpus (Kwon et al., 2022) consists of binary jargon annotation for sentences in the electronic health record (EHR) notes, whereas the dataset is licensed. Other work on complex word identification mainly focuses on general domains, such as news and Wikipedia, and other specialized domains, e.g., computer science. Due to space limits, we list them in Appendix E. Our data is based on open-access medical resources and contains both sentence-level readability ratings and complex span annotation with a finer-grained 7-class categorization (§2).

7 Conclusion

In this work, we present a systematic study for sentence readability in the medical domain, featuring a new annotated dataset and a data-driven study to answer *"why medical sentences are so hard."*. In the analysis, we quantitatively measure the impact of several key factors that contribute to the complexity of medical texts, such as the use of jargon, text length, and complex syntactic structures. Future work could extend to the medical notes from clinical settings to better understand real-time communication challenges in healthcare. Additionally, leveraging our dataset that categorizes complex spans by difficulty and type, further research could develop personalized simplification tools to adapt content to the target audience, thereby improving patients' understanding of medical information.

Limitations

Due to the reality that major scientific medical discoveries are mostly 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 focus of our work is to create readability datasets for general purposes following prior work. We did not study or distinguish the fine-grained differences and nuances between native speakers and non-native speakers (Yimam et al., 2017).

The readability ratings of a sentence can be impacted by a mixture of factors, including sentence length, grammatical complexity, word difficulty, the 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 widely 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*).

Ethics Statement

During the data collection process, we hired undergrad students from the U.S. as in-house annotators. All annotators are compensated at $18 per hour or by credit hours based on the university standards.

Acknowledgments

The authors would like to thank Mithun Subhash, Jeongrok Yu, and Vishnu Suresh for their help in data annotation. This research is supported in part by the NSF CAREER Award IIS-2144493, NSF Award IIS-2112633, NIH Award R01LM014600, ODNI and IARPA via the HIATUS program (contract 2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, NIH, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

AI@Meta. 2024. Llama 3 model card.

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, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

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

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

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, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020a. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020b. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

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.

Marc Brysbaert and Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand english word meanings. *Behavior research methods*, 49:1520–1523.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44:991–997.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

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, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021b. 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.

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 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, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Wang, and Tal August. 2024. Personalized jargon identification for enhanced interdisciplinary communication. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4535–4550, Mexico City, Mexico. 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*, abs/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.

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Understanding jargon: Combining extraction and generation for definition modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 Jessy 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.

David Kauchak and Gondy Leroy. 2016. Moving beyond readability metrics for health-related text simplification. *IT professional*, 18(3):45–51.

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.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

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.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. 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*, abs/1907.11692.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

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.

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*, abs/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, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 851–860. ACM.

David M. W. Powers. 1998. Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*.

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.

Daniel Simig, Tianlu Wang, Verna Dankers, Peter Henderson, Khuyagbaatar Batsuren, Dieuwke Hupkes, and Mona Diab. 2022. Text characterization toolkit (TCT). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 72–87, Taipei, Taiwan. Association for Computational Linguistics.

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

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4096–4102. ijcai.org.

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.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *ArXiv preprint*, abs/1603.06009.

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.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

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:

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

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:

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

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, S is sentence length, 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 More Results on the Influence of Each Linguistic Feature

In this section, we provide more results on the influence of linguistic features, including syntax and semantics features, quantitative and corpus linguistics features, in addition to psycho-linguistic features (Vajjala and Meurers, 2016), such as the age of acquisition (AoA) released by Kuperman et al. (2012), and concreteness, meaningfulness, and imageability extracted from the MRC psycholinguistic database (Wilson, 1988).

The features are extracted using a combination of toolkits, each of which covers a different subset of features, including 220 features from the `LFTK` package (Lee and Lee, 2023), 255 from the `LingFeat` (Lee et al., 2021), 61 from `Text Characterization Toolkit (TCT)` (Simig et al., 2022), 119 from `Profiling-UD` (Brunato et al., 2020a), 33 from the `Lexical Complexity Analyzer (LCA)` (Lu, 2012) and 23 from the `L2 Syntactic Complexity Analyzer (L2SCA)` (Lu, 2010). The top 50 most influential features are presented in Table B after skipping the duplicated and nearly equivalent ones, e.g., the *typo-token-ratio* and *root-type-token-ratio*.

For each of the listed features, we look into the implementation details from the original toolkit and explain them in the "Implementation Details" column. To facilitate reproducibility, we also include the exact feature name used in the original code in the "Original Feature Name" column.

Package	Original Feature Name	Pearson Correlation	Implementation Details in the Original Toolkit
LCA (2012)	`len(slextypes.keys())`	0.6452	Number of unique sophisticated lexical words, which are lexical words (i.e., nouns, non-auxiliary verbs, adjectives, and certain adverbs that provide substantive content in the text) and are also "sophisticated" (i.e., not in the list of 2,000 most frequent lemmatized tokens in the ANC[a] corpus).
LCA (2012)	`len(swordtypes.keys())`	0.6408	Number of unique sophisticated words. "Sophisticated" is defined as not in the list of 2,000 most frequent lemmatized tokens in the `American National Corpus` (ANC)
LFTK (2023)	`corr_ttr`	0.6271	Corrected type-token-ratio (CTTR), which is calculated as $(\text{number-of-unique-tokens}/\sqrt{2 \times \text{number-of-all-tokens}})$, based on the lemmatized tokens.
LFTK (2023)	`corr_ttr_no_lem`	0.6158	Corrected type-token-ratio (CTTR), which is calculated as $(\text{number-of-unique-tokens}/\sqrt{2 \times \text{number-of-all-tokens}})$, based on the tokens without lemmatization.
LCA (2012)	`slextokens`	0.6120	Number of all sophisticated lexical words, which are lexical words (i.e., nouns, non-auxiliary verbs, adjectives, and certain adverbs that provide substantive content in the text) and are also "sophisticated" (i.e., not in the list of 2,000 most frequent lemmatized tokens in the ANC corpus).
LCA (2012)	`swordtokens`	0.6083	Number of all sophisticated words. "Sophisticated" is defined as not in the 2,000 most frequent lemmatized tokens in the `American National Corpus` (ANC)
LCA (2012)	`ndwz`	0.6037	Number of different words in the first Z words. Z is computed as the 20th percentile of word counts from a dataset, resulting in a value of 16 in our case.
LCA (2012)	`ndwesz`	0.6024	Number of different words in expected random sequences of Z words over ten trials. Z is computed as the 20th percentile of word counts from a dataset, resulting in a value of 16 in our case.
LingFeat (2021)	`WRich20_S`	0.6006	Semantic richness of a text, which is calculated by summing up the probabilities of 200 Wikipedia-extracted topics, each multiplied by its rank, indicating the text's variety and depth of topics. The 200 topics were extracted from the Wikipedia corpus using the Latent Dirichlet Allocation (LDA) method.
LCA (2012)	`len(lextypes.keys())`	0.5996	Number of unique lexical words. Lexical words include nouns, non-auxiliary verbs, adjectives, and certain adverbs that provide substantive content in the text.
LCA (2012)	`ndwerz`	0.5961	Number of different words expected in random Z words over ten trials. Z is computed as the 20th percentile of word counts from a dataset, resulting in a value of 16 in our case.
LFTK (2023)	`t_syll`	0.5888	Number of syllables.
LFTK (2023)	`t_char`	0.5806	Number of characters.
TCT (2022)	`WORD_PROPERTY_AOA_MAX`	0.5758	Max age-of-acquisition (AoA) of words. The AoA of each word is defined by Kuperman et al. (2012).
LCA (2012)	`lextokens`	0.5750	Number of lexical words. Lexical words include nouns, non-auxiliary verbs, adjectives, and certain adverbs that provide substantive content in the text.

Table 10: Top 50 most influential linguistic features on readability assessment.

[a] `https://anc.org/`

Package	Original Feature Name	Pearson Correlation	Implementation Details in the Original Toolkit
LFTK (2023)	t_uword	0.5744	Number of unique words.
LingFeat (2021)	WTopc20_S	0.5686	The count of distinct topics, out of 200 extracted from Wikipedia, that are significantly represented in a text, showing the breadth of topics it covers.
LFTK (2023)	t_syll2	0.5607	Number of words that have more than two syllables.
LingFeat (2021)	BClar20_S	0.5598	Semantic Clarity measured by averaging the differences between the primary topic's probability and that of each subsequent topic, reflecting how prominently a text focuses on its main topic, based on 200 topics extracted from the WeeBit Corpus.
LingFeat (2021)	to_AAKuW_C	0.5379	Total age-of-acquisition (AoA) of words. The AoA of each word is defined by Kuperman et al. (2012).
TCT (2022)	DESWC	0.5323	Number of words.
LingFeat (2021)	BClar15_S	0.5294	Semantic Clarity measured by averaging the differences between the primary topic's probability and that of each subsequent topic, reflecting how prominently a text focuses on its main topic, based on 150 topics extracted from the WeeBit Corpus.
LingFeat (2021)	at_Chara_C	0.5237	Average number of characters per token.
LFTK (2023)	corr_noun_var	0.5127	Corrected noun variation, which is computed as $(\text{number-of-unique-nouns}/\sqrt{2 \times \text{number-of-all-nouns}})$
LingFeat (2021)	as_AAKuW_C	0.5069	Average age-of-acquisition (AoA) of words. The AoA of each word is defined by Kuperman et al. (2012).
LFTK (2023)	t_bry	0.5046	Total age-of-acquisition (AoA) of words. The AoA of each word is defined by Brysbaert and Biemiller (2017).
LFTK (2023)	t_syll3	0.5044	Number of words that have more than three syllables.
LingFeat (2021)	WTopc15_S	0.4956	The count of distinct topics, out of 150 extracted from Wikipedia, that are significantly represented in a text, showing the breadth of topics it covers.
LFTK (2023)	corr_adj_var	0.4764	Corrected adjective variation, which is computed as $(\frac{\text{number-of-unique-adjectives}}{\sqrt{2 \times \text{number-of-all-adjectives}}})$
LFTK (2023)	n_unoun	0.4694	Number of unique nouns.
LingFeat (2021)	at_Sylla_C	0.4691	Average number of syllables per token.
LFTK (2023)	a_bry_ps	0.4586	Average age-of-acquisition (AoA) of words. The AoA of each word is defined by Brysbaert and Biemiller (2017).
LFTK (2023)	n_noun	0.4581	Number of nouns.
LingFeat (2021)	to_FuncW_C	0.4515	Number of function words, excluding words with POS tags of 'NOUN', 'VERB', 'NUM', 'ADJ', or 'ADV'.
LFTK (2023)	n_adj	0.4497	Number of adjectives.
LFTK (2023)	n_uadj	0.4483	Number of unique adjectives.
Profiling–UD (2020b)	avg_max_depth	0.4371	The maximum tree depths extracted from a sentence, which is calculated as the longest path (in terms of occurring dependency links) from the root of the dependency tree to some leaf.
LingFeat (2021)	WNois20_S	0.4362	Semantic noise, which quantifies the dispersion of a text's topics, reflecting how spread out its content is across different subjects. It is calculated by analyzing the text's topic probabilities on 200 topics extracted from through Latent Dirichlet Allocation (LDA).
LCA (2012)	ls1	0.4255	Lexical Sophistication-I, calculated as the ratio of sophisticated lexical tokens to the total number of lexical tokens.

Table 11: Top 50 most influential linguistic features on readability assessment (continue).

Package	Original Feature Name	Pearson Correlation	Implementation Details in the Original Toolkit
LFTK (2023)	t_subtlex_us_zipf	0.4253	Cumulative Zipf score for all words, based on frequency data from the SUBTLEX-US corpus (Brysbaert et al., 2012). Zipf scores are a measure of word frequency, with higher scores indicating more common words.
LingFeat (2021)	WTopc10_S	0.4242	The count of distinct topics, out of 100 extracted from Wikipedia, that are significantly represented in a text, showing the breadth of topics it covers.
Profiling–UD (2020b)	avg_links_len	0.4167	Average number of words occurring linearly between each syntactic head and its dependent (excluding punctuation dependencies).
LFTK (2023)	n_adp	0.4144	Number of adpositions.
LingFeat (2021)	SquaAjV_S	0.4088	Squared Adjective Variation-1, which is calculated as the $(\frac{(\text{number-of-unique-adjectives})^2}{\text{number-of-total-adjectives}})$.
LFTK (2023)	n_upunct	0.4053	Number of unique punctuations.
LFTK (2023)	corr_adp_var	0.4031	Corrected adposition variation, which is computed as $(\frac{\text{number-of-unique-adpositions}}{\sqrt{2 \times \text{number-of-all-adpositions}}})$
LFTK (2023)	n_uadp	0.4022	Number of unique adpositions.
LFTK (2023)	corr_propn_var	0.3895	Corrected proper noun variation, which is computed as $(\frac{\text{number-of-unique-proper-nouns}}{\sqrt{2 \times \text{number-of-all-proper-nouns}}})$
LingFeat (2021)	WClar20_S	0.3879	Semantic Clarity measured by averaging the differences between the primary topic's probability and that of each subsequent topic, reflecting how prominently a text focuses on its main topic, based on 200 topics extracted from Wikipedia Corpus.
LingFeat (2021)	SquaNoV_S	0.3864	Squared Noun Variation-1, which is calculated as the $((\text{number-of-unique-nouns})^2/\text{number-of-total-nouns})$.

Table 12: Top 50 most influential linguistic features on readability assessment (continue).

C Introduction of Medical Text Simplification Resources

Our dataset is constructed on top of open-accessed resources. Each of the resources is detailed below. Table 13 presents the basic statistics of 180 sampled article (segment) pairs.

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, 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.

[9] https://www.cochranelibrary.com/

[10] https://query.wikidata.org/

[11] The March 22, 2023 version.

[12] https://attardi.github.io/wikiextractor/

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

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

D 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.

E More Related work on Complex Span Identification in Medical Domain

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). In addition, Guo et al. (2024) collects a jargon dataset from computer science research papers, Lucy et al. (2023) studies the social implications of jargon usage, and August et al. (2022); Huang et al. (2022) focus on the explanation of jargon.

[13] https://github.com/huggingface/transformers

F More Results for Complex Span Identification

Table 14 presents the results of the exact match at entity level for the complex span identification task on the MEDREADME test set. 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)	<u>73.5</u>	<u>70.0</u>	<u>62.4</u>
BioBERT (2020)	70.5	67.1	59.8
PubMedBERT (2021)	72.2	69.0	61.2

Table 14: **Micro F1** of exact match at entity-level for complex span identification task on the MEDREADME test set. The **best** and <u>second best</u> scores within each model size are highlighted. Models are trained with fine-grained labels in seven categories and evaluated at different granularity.

G More Results on Medical Readability Prediction

We conducted an additional experiment to study how different complex span identification models used in Section 5 affect the performance of medical readability prediction. We find that using predictions from different complex span prediction models leads to similar improvements in readability prediction, with a $\pm$ 0.015 difference in average Pearson correlation across different resources.

H 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 15: 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.

I Annotated Screenshot of Search Engine Results

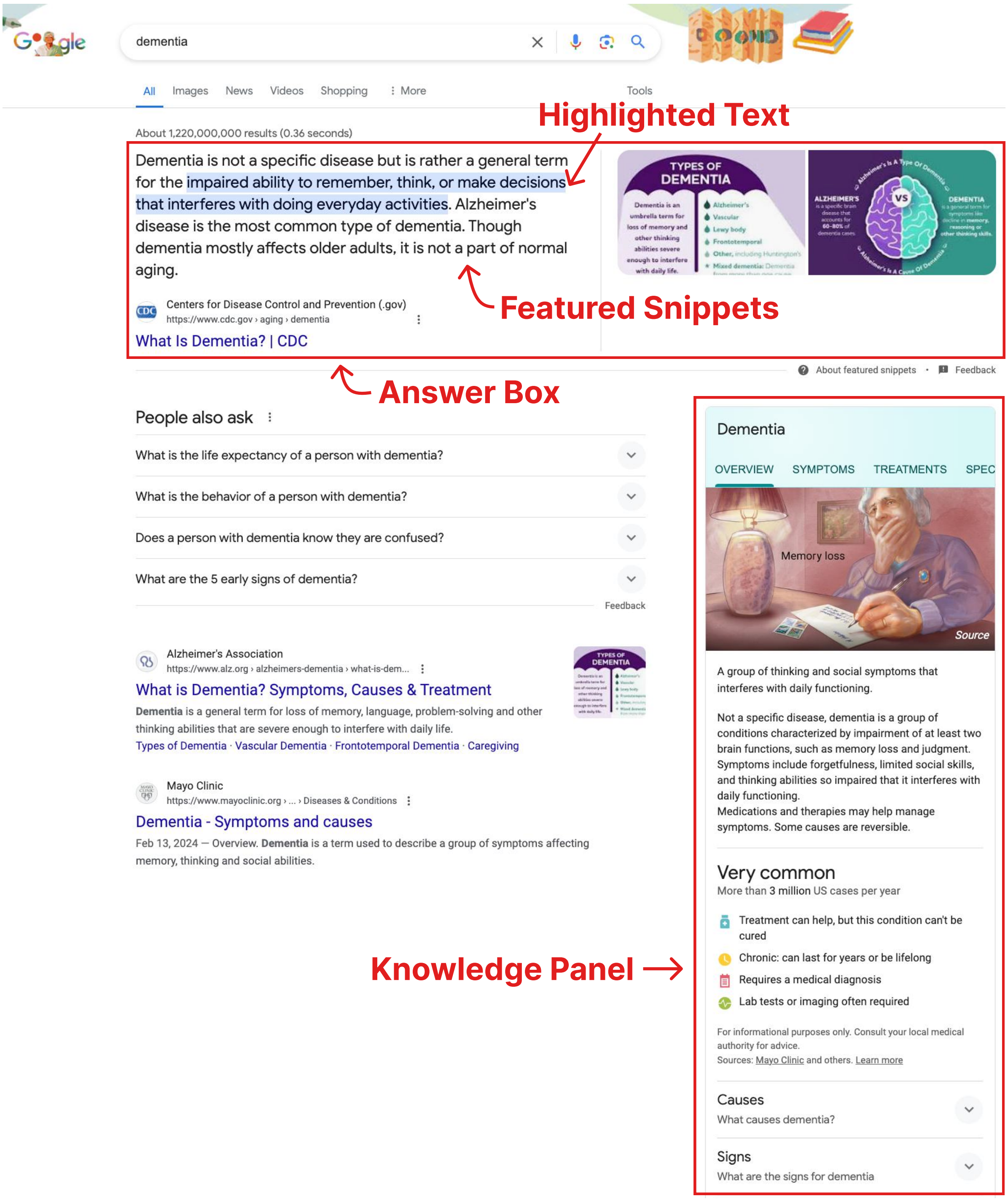
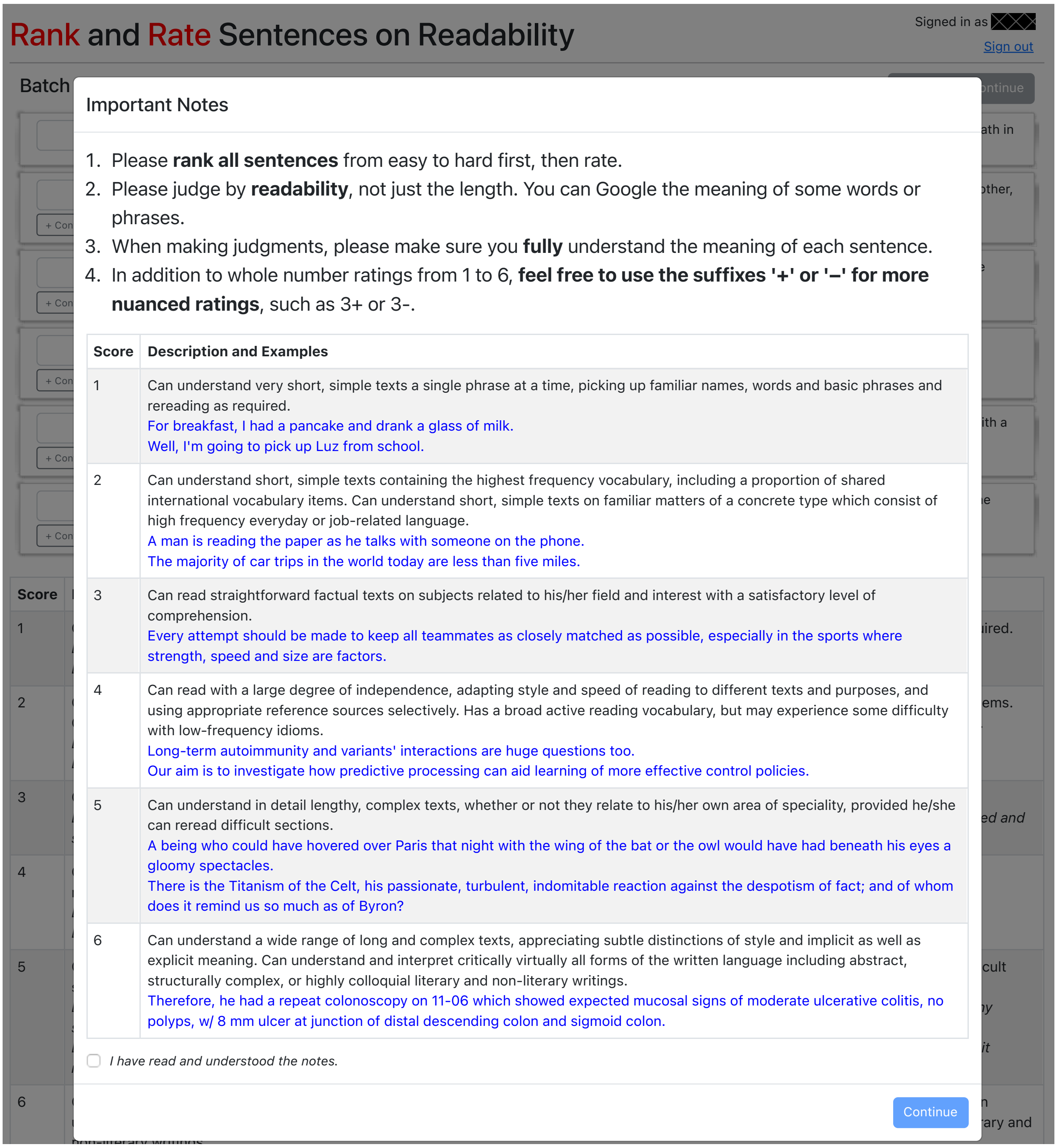


Figure 6: An annotated screenshot of search results from Google. Search engines may provide the explanation of a medical term in two places: (1) the feature snippets in the answer box and (2) the knowledge panel on the right-hand side, which is powered by a knowledge graph.

J Annotation Interface for Sentence Readability



Rank and Rate Sentences on Readability

Signed in as [redacted]

Sign out

Important Notes

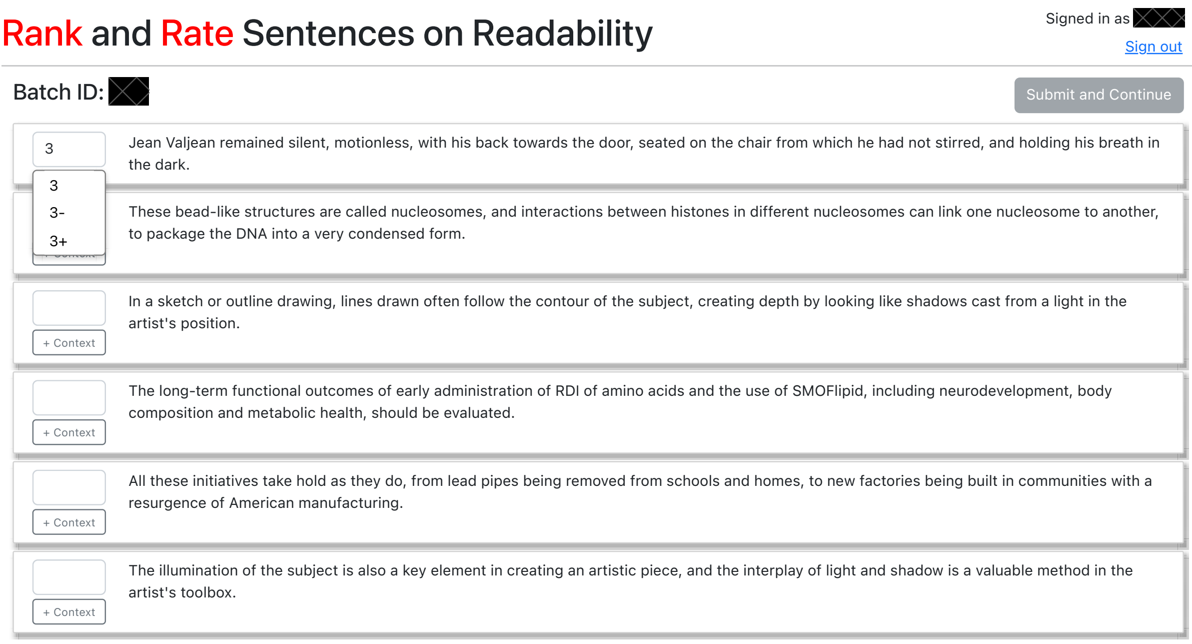
1. Please **rank all sentences** from easy to hard first, then rate.
2. Please judge by **readability**, not just the length. You can Google the meaning of some words or phrases.
3. When making judgments, please make sure you **fully** understand the meaning of each sentence.
4. In addition to whole number ratings from 1 to 6, **feel free to use the suffixes '+' or '–' for more nuanced ratings**, such as 3+ or 3-.

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. For breakfast, I had a pancake and drank a glass of milk. 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. A man is reading the paper as he talks with someone on the phone. 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. 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. Long-term autoimmunity and variants' interactions are huge questions too. 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. 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. 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. 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.

I have read and understood the notes.

Continue

Figure 7: 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 8: The interface for annotating sentence readability. Annotators can click the "+ Context" button to see the surrounding sentences.

K Annotation Interface for Complex Span Identification

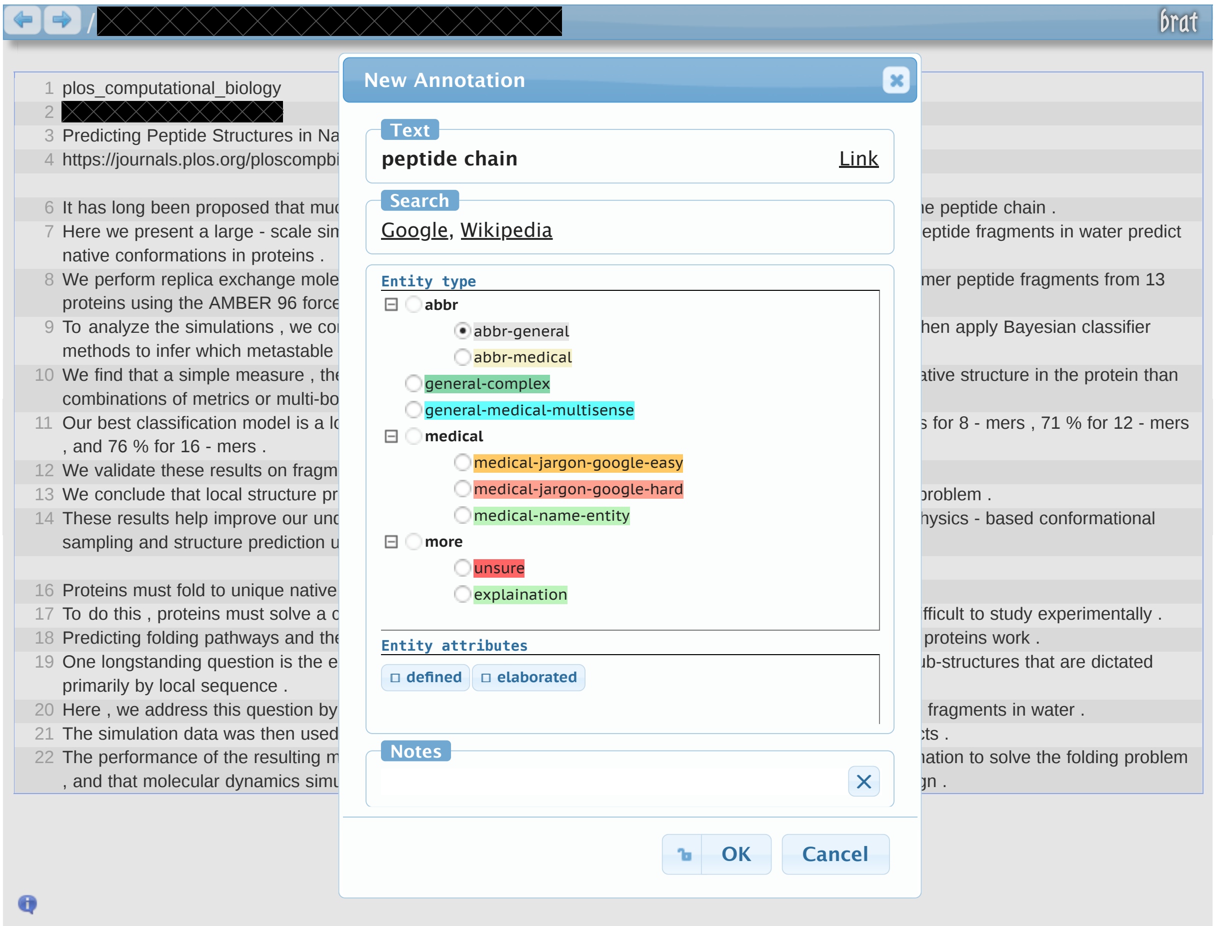


Figure 9: The annotation interface for complex span identification.

L Annotation Guideline for Complex Span Identification

Assumption:

- English-speaking (both native and non-native) with a college-level education background. With access to Google and dictionaries.

TL;DR:

- Abbr-general
- Abbr-medical
- General-complex
- General-Medical-multisense (same word, different meaning under medical context)
- Medical-jargon-google-easy (medical terms, but can be easily looked up in Google or Medical Dictionary)
- Medical-jargon-google-hard (often multi-word expression which is hard to understand even with Google, even if each individual word might be easy to understand)
- Medical-name-entity (something like a specific brand name, “Pfizer,” not general medicine or equipment name)

Guidelines:

- **Abbreviation:** (new Rule)
  - **[Abbr-medical]** Abbreviation with a specific meaning in the medical domain.
    - TB: tuberculosis
    - BP: blood pressure
    - BID: twice a day
  - **[Abbr-general]** an abbreviation that does *NOT* belong to the medical domain, including statistical terms
    - MD: mean deviation,
    - RCT: randomized controlled trial
    - CI: confidence intervel

- **[General-complex]** **General complex words/phrases**: words that are outside the vocabulary of 10-12th graders and *NOT* specific/strictly to the medical domain
    - aberrant
    - tender
    - ammonia
    - Rule of thumb: Assuming you are a normal college student without a medical background, can you understand the term? If not, put it here.

- **[General-Medical-multisense]** **Multi-sense terms**: layman terms that have specific meanings in the clinical context, which are off from their general meaning. *[Note: This might be the hardest category for lay annotators to annotate.]*
    - accommodate: a drug or substance that stops the action or effect of another substance.
    - antagonize: when the eye changes focus from far to near
    - formed: stool that is solid
    - resident: a physician receiving specialized clinical training in a hospital

Figure 10: The annotation guideline for complex span identification.

- **Medical complex words/phrases**: medical terminology that may be unfamiliar to persons without clinical experience (Note: For medical terms, if you don't know the 99% nearly "exact meaning," please annotate it; "roughly getting it" is not enough.)

  - **[Medical-jargon-google-easy] Medical/technical terms** that are (1) outside the vocabulary of 10-12th graders, (2) specific to the medical domain, (3) but can easily be understood by Google or Dictionary.
    - tonsils: a part of the body.
    - airway protection: insert a tube from the outside to the inside to open up the upper airways for the patient
    - monotherapy

  - **[Medical-jargon-google-hard]** Often multi-word expressions. They mean something different or are difficult to understand quickly from their individual parts.
    - hazard of a disease-free survival event
    - treatment-by-time interaction
    - Rule of thumb: after Googling, can you confidently explain this to another person? If not, put it here.

  - **[Medical-name-entity]** For example, the brand name. But please exclude general medical terms, such as medicine and medical equipment.
    - Pfizer, Moderna: company name
    - Cochrane, Embase: medical publication platform
    - But not: norfloxacin, artificial cardiac pacemaker

Additional Notes:

- Medical terms whose definitions are widely known do NOT need to be labeled. (e.g., muscle)
- On top of the annotation of medical terms following the categorizations below, annotators can flag an *optional* attribute to reflect if the medical term is elaborated or defined/explained in the content.
  - Elaborated: carbon dioxide, which is the gas we breathe out.
    - Here, carbon dioxide is not defined but just further elaborated with an example.
  - Defined: Hazard ratio (HR).
    - Here, HR is defined as the hazard ratio.

Figure 11: The annotation guideline for complex span identification (continue).