PlainMedScale: A Corpus of Multi-Level Simplified Medical Texts in German and English

Bruno Brocai
Heidelberg University
German Linguistics

Ilaria Papagno
Heidelberg University
Translation Studies

Mayumi Ohta
Heidelberg University
SFB 1671

{bruno.brocai@gs, ilaria.papagno@iued, ohta@cl}.uni-heidelberg.de

Abstract

We introduce **PlainMedScale**, a topic-aligned medical corpus spanning four levels of comprehensibility in German and English, drawn from *MSD Manuals* (professional and consumer), *Gesund.Bund*, *Apotheken Umschau* (*Einfache Sprache*), and the *NHS*. The four tiers correspond to distinct communicative functions — reference, explanation, decision support, and access — and move beyond the binary expert–lay contrast of prior corpora. In two pilot studies enabled by the alignments, we show that many readability metrics established on two registers fail to generalize across the full gradient, and that a SOTA open-weight LLM prompted for Plain Language still partially preserves the difficulty of its input. Code (https://github.com/GS-Uni-Heidelberg/PlainMedScale) and data (10.5281/zenodo.21728290) are made available.

1 Introduction

Providing high-quality, accessible medical information has been widely acknowledged as a key determinant of enhancing health literacy, defined as individuals' capability to locate, comprehend, and utilize health-related information in their daily lives (Sørensen et al., 2012; Council of Europe, 2023). Simplification of medical texts into Plain Language should provide access to relevant medical information for a wider audience, including citizens with low literacy skills, cognitive impairments, or a second language background (Inclusion Europe, n.d.; Laban et al., 2021). Recent work on Plain Language (*Einfache Sprache*) guidelines (Maaß, 2020; Bock, 2015) supports the translation of specialized content into lay-friendly texts. At the same time, automatic text simplification has advanced considerably (Lyu and Pergola, 2024), offering a potential solution for making specialized knowledge more accessible at scale — provided suitable corpora.

However, existing German medical text corpora exhibit some limitations. Many present sentence alignment, which enables a detailed analysis of linguistic transformations, but fail to capture content transposition, tending to produce same-length document pairs, without considering the summarization nature of the simplification process, as pointed out by Aumiller and Gertz (2022). Most of the existing sources align two registers, a professional version and a single simplified counterpart, thereby reducing medical text simplification to a binary expert-lay contrast. Yet, medical text simplification is not a uniform process but rather a continuum spanning multiple degrees of accessibility. Capturing this gradient requires aligning more than two levels of difficulty. To this end, we extend the English and German texts of the multilingual multiMSDCorpus (Horiguchi et al., 2025), with additional lay-oriented resources in both languages. The resulting document-aligned corpus integrates four complementary sources per language, covering different degrees of medical language difficulty. This corpus serves two key purposes: (1) to demonstrate, through a comparison of different steps of linguistic and content transformations across registers, that conventional patient versions no longer meet the required level of acceptability; and (2) to provide four aligned sources at different difficulty levels, offering a more comprehensive picture of the many facets of medical text simplification.

2 Related Work

For English, Med-EASi (Basu et al., 2023) provides short expert–lay text pairs annotated with simplification strategies (elaboration, replacement, insertion, deletion), while JEBS (Xia et al., 2025) is a larger resource targeting fine-grained lexical simplification of complex biomedical terms in scientific abstracts. At a coarser granularity, the Cochrane corpus (Devaraj et al., 2021) aligns tech-

Source	Text
MSD Prof.	*Etiology of Appendicitis* — Appendicitis is thought to result from obstruction of the appendiceal lumen, typically by lymphoid hyperplasia but occasionally by a fecalith, foreign body, tumor, or even worms. The obstruction leads to distention, bacterial overgrowth, ischemia, and inflammation. If untreated, necrosis, gangrene, and perforation occur. If the perforation is contained by the omentum, an appendiceal abscess results.
MSD Consumer	*Causes of Appendicitis* — The cause of appendicitis is not fully understood. However, in most cases, a blockage inside the appendix probably starts a process. The blockage may be from a small, hard piece of stool (fecalith), a foreign body, tumor, or, rarely, even worms. As a result of the blockage, the appendix becomes inflamed and infected. If inflammation continues without treatment, the appendix can rupture.
Gesund.Bund	*What causes appendicitis?* — The appendix is a small protuberance of the cecum that has a narrow lumen (opening). If this lumen is bent or blocked, the appendix swells, interrupting the blood flow. The cells in the wall of the appendix die off and inflammation develops, fostered by bacteria from the intestine. The often purulent inflammation attacks the wall of the appendix, potentially creating a hole.
NHS	*Causes of appendicitis* — The appendix is a small pouch that's joined to your bowel in the lower right side of your abdomen (tummy). Appendicitis happens when your appendix becomes infected and swollen. This is often caused by something getting stuck in your appendix, such as a small piece of undigested food or hard poo.

Table 1: Parallel passages on the etiology of appendicitis across the four sources, illustrating increasing degrees of simplification.

nical abstracts with plain-language summaries at the paragraph level, and MultiCochrane (Joseph et al., 2023) extends it to sentence-aligned pairs in English, Spanish, French, and Farsi.

Comparable efforts exist for German, though some remain inaccessible due to data protection restrictions (Klaper et al., 2013; Borchert et al., 2022). General-purpose resources include Battisti et al. (2020)'s multi-domain corpus for automatic readability assessment and text simplification spanning 92 domains, Jablotschkin et al. (2024)'s DE-Lite, which integrates existing corpora with new online data, and Aumiller and Gertz (2022)'s document-aligned corpus derived from the children's encyclopedia Klexikon. Further corpora include DEplain (Stodden et al., 2023), which provides document- and sentence-level resources outside the medical domain, and the sentence-aligned dataset of Toborek et al. (2023). Within medicine, the multilingual multiMSDCorpus (Horiguchi et al., 2025) aligns text across nine languages, including German. However, its patient-oriented version still shows limited readability and a high density of specialized terminology. We therefore propose integrating these text pairs with additional patient-oriented resources offering higher readability and official plain-language resources.

3 Dataset

3.1 Sources

MSD (MSD Manuals, 2025) is a widely used medical resource offering articles in two versions, one for professionals and one for patients, of which we use the German and English editions. *Gesund.Bund* (Gesund.Bund, 2025), provided by the German Federal Ministry of Health, presents conditions and health-related topics in everyday language with little to no technical vocabulary; we again select the German and English editions. The *Einfache Sprache* section of *Apotheken Umschau* (Apotheken Umschau, 2020) comprises certified plain language texts on diseases and health-related topics in German. At a comparable level in English are the condition descriptions from the official *NHS* website (NHS, 2025), which targets a reading age of 9 to 11. From these sources, the condition sections (but not, e.g., the medication sections) were scraped.

	Source (desc. difficulty)	N	Aligned w/ k tiers: 1	2	3
German	*MSD Prof.*	1594	1360	157	77
	MSD Consumer	1594	1360	157	77
	Gesund.Bund	354	42	108	65
	ApoUm	180	30	28	81
English	*MSD Prof.*	1594	1286	197	111
	MSD Consumer	1594	1286	197	111
	Gesund.Bund	354	8	126	47
	NHS	500	9	166	48

Table 2: Corpus composition. N = articles per source; columns 1/2/3 = articles aligned with that many of the other three tiers.

Model (reasoning)	Overall	Per-class accuracy (%) corr.	spec./gen.	inc.
GPT-5.4-mini (none)	80.0	77.1	41.4 / 66.7	90.2
GPT-5.4-mini (low)	80.1	71.8	60.9 / 75.0	88.1
GPT-5.4 (none)	80.2	69.4	64.8 / 75.0	88.4
Claude Haiku 4.5 (none)	77.3	73.9	48.4 / 66.7	85.5
Qwen3-30B-Instruct (-)	75.3	88.2	21.9 / 8.3	83.2
n	956	245	128 / 12	571

Table 3: Four-class LLM-judge accuracy on the 956-pair manual gold set.

3.2 Instance Alignment

Articles from sources with parallel variants (*Gesund.Bund* across languages, *MSD Manuals* across difficulty levels) are aligned via site metadata. The remaining articles are routed through shared medical reference vocabularies — ICD-10 (Bundesinstitut für Arzneimittel und Medizinprodukte, 2024), MeSH (U.S. National Library of Medicine, 2025), German MeSH (Fischer-Wagener et al., 2023), the Human Disease Ontology (DO) (Baron et al., 2026), and Pschyrembel (Pschyrembel, 2025), and a pair is considered aligned when both articles resolve to the same concept at least once.

Stage 1: Lexical alignment. ICD-10 codes embedded in source HTML are extracted at scrape time. To broaden coverage, we index every dictionary entry and its synonyms in SQLite, lemmatized with spaCy (Honnibal et al., 2020), and look up each article's canonical title and aliases against the matching-language dictionaries. Matched identifiers propagate across vocabularies via DO cross-references (DOID $\leftrightarrow$ MeSH $\leftrightarrow$ ICD-10CM).

Stage 2: Embedding-based alignment. Lexical matching is high-recall but noisy. We embed each article's first paragraph and every dictionary entry with Qwen3-Embedding-8B (Zhang et al., 2025) and (i) re-rank Stage 1 candidates by L2 distance, retaining the closest; (ii) for articles without any Stage 1 hit, run top-k retrieval against each dictionary, accepting a match only if a reverse full-text lookup of its title resolves back to the same article. Two articles are linked whenever a chain of shared concept IDs connects them, and we rank such links by the embedding distance between the articles.

Stage 3: LLM-based filtering. Two annotators with medical backgrounds labeled Apotheken-Umschau-anchored pairs as `correct` (the same topic), `specialization` (source is narrower than *Apotheken Umschau*), `generalization` (broader), and `incorrect` (unrelated), resulting in 956 pairs. The annotation showed that many pairs are partial topic matches (e.g., *Paralysis–Vocal Fold Paralysis*), and that Stage 1+2 precision is low overall.

We evaluate different candidate models (Table 3) to be used in the follow-up filtering step. We observe a tradeoff: larger or reasoning-enabled models do better on the harder `specialization/generalization` distinction, but this comes at the cost of accuracy on the `correct/incorrect` classes — the two that matter most for corpus quality. GPT-5.4-mini flags the most genuine `incorrect` alignments while trailing only Qwen3 on `correct`; Qwen3's advantage, however, comes with weak `incorrect` detection. We therefore pick GPT-5.4-mini as our judge for the full data.

We noticed the original labels sometimes collapsed a cluster of related titles into one judgment, so a pair that was actually a distinct match could inherit a label that did not fit. To quantify this, a second reviewer re-annotated 100 pairs, yielding 92 % precision on `correct` (95 % Wilson CI 67–99) and $\geq$ 88 % on the other classes. Both numbers are biased: the 92 % likely overestimates true agreement, since the reviewer saw each LLM label and was asked to override only disagreements (anchoring bias toward the LLM's judgment); the original reviewer accuracy likely underestimates it, due to the issue described above. True precision most likely falls between the two, a level we consider acceptable.

4 Text Assessment

4.1 Readability Indices

We apply a subset of the readability metrics from Scholz and Wenzel (2025), evaluated there on German and English Plain Language summaries; we refer to that paper for the exact equations and NLP pipeline. We report two additional metrics: since Flesch-Kincaid Grade Level is not suited for German, we report the Wiener Sachtextformel (Bamberger and Vanecek, 1984) for that language, and we estimate the frequency of medical jargon via a corpus-based keyness approach (full procedure in appendix A). Table 4 reports the German results (English in appendix B.3).

The aligned nature of our corpus lets us test the statistical significance of readability metrics

Metric	Apotheken Umschau		Gesund.Bund		MSD Cons.		MSD Prof.
Wiener Sachtextformel	6.80 ± 1.05	*	10.29 ± 1.03	*	10.76 ± 1.24	*	12.23 ± 1.52
Noun ratio	0.30 ± 0.02	†	0.27 ± 0.02	*	0.33 ± 0.03	*	0.36 ± 0.05
Adjective ratio	0.04 ± 0.02	*	0.06 ± 0.02		0.07 ± 0.02	*	0.09 ± 0.02
Function word ratio	0.68 ± 0.02	†	0.69 ± 0.02	*	0.68 ± 0.03	†	0.71 ± 0.04
Numbers ratio	0.00 ± 0.01	*	0.01 ± 0.00		0.01 ± 0.01	*	0.02 ± 0.02
Negations ratio	0.01 ± 0.00		0.01 ± 0.00	*	0.01 ± 0.00	*	0.01 ± 0.00
Grammar frequency	0.85 ± 0.05	†	0.95 ± 0.04	*	0.89 ± 0.08	*	0.87 ± 0.12
Term frequency (PMW)	70.3 ± 15.3	*	54.1 ± 10.4	*	44.3 ± 12.0	*	34.4 ± 11.5
Avg. lexical chain length	0.01 ± 0.00	*	0.00 ± 0.00	*	0.01 ± 0.00	†	0.01 ± 0.01
Avg. lexical cross chains	0.49 ± 0.18	*	0.82 ± 0.32	*	1.00 ± 0.81	†	0.89 ± 0.83
Jargon density	0.043 ± 0.022	*	0.074 ± 0.025	*	0.099 ± 0.031	*	0.148 ± 0.041

Table 4: Readability metrics (DE) positively evaluated by Scholz and Wenzel (2025), mean ± std across articles per source. Markers between adjacent low→high-readability data sources indicate the significant shift in the expected direction (*), or in the opposite direction (†) (one-sided Wilcoxon signed-rank on aligned article pairs, α=0.05, Bonferroni-corrected). Because the test was only done on aligned pairs, full-corpus mean and significance might not agree. Bold/underlined: metrics that are fully/partially significant only in the expected direction.

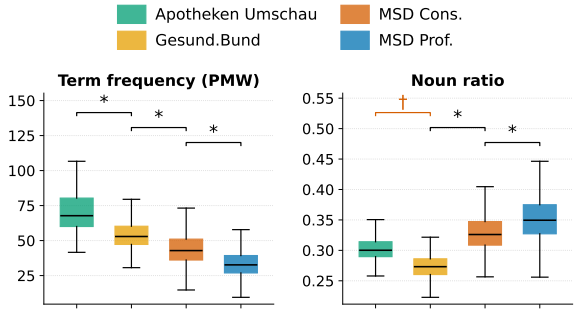

Figure 1: Two metrics from Table 4: term frequency shifts as expected at every step, noun ratio does not (†).

across difficulty levels while minimizing the topic confounder, using a Wilcoxon signed-rank test on each pairwise difficulty step. Surprisingly, most readability metrics show no clear slope across the four sources: the noun ratio, for instance, is lowest for *Gesund.Bund* and significantly higher for *Apotheken Umschau*. The only metrics that hold up fully for German are the Wiener Sachtextformel, the adjective and negation ratios, and the related term Frequency and jargon density measures.

Overall, these findings point to a problem in evaluating readability metrics on only two corpora: any score difference observed there may hold at that specific difficulty step without generalizing across the full readability range.

4.2 Content Analysis

The sources in our dataset differ not only in linguistic complexity, but also in their communicative function (cf. Jakobson, 1960). Through a qualitative reading of a sample of aligned articles, we identified four shared functions, each of which was more prominent in a single tier: a reference function (*MSD Prof.*), presenting information flatly and systematically; an explanatory function (*MSD Consumer*), conveying pathophysiological accounts to lay readers through narrative framing; a decision-support function (*Gesund.Bund*), providing evidence-based statements and quality-of-life considerations to inform autonomous patient decisions; and an access function (*Apotheken Umschau* Plain Language, *NHS*), offering action-oriented, low-threshold information organized around readers' likely concerns. This functional differentiation explains why certain information appears in some sources but not in others, and provides a normative criterion for what should be preserved during simplification.

Plain language must not only remove linguistic obstacles — such as complex syntax and jargon — but also reshape content. Defining a threshold of content completeness is challenging, since omission is intrinsic to simplification and its adequacy is determined by its function. Comparing different levels of human simplification helps identify what should be retained. Patient-oriented versions of MSD texts still contain substantial medical jargon — uncommon anatomical terms, adjectives, and verbs — often left undefined or explained only later in the text, while illustrations rely on classical terminology. Their communicative function remains referential: information is presented without a clear hierarchy of relevance and offers little practical guidance. *Gesund.Bund* and *Apotheken Umschau*, by contrast, indicate when a symptom requires medical intervention or when a given procedure is routine, guiding readers' judgment without demand-

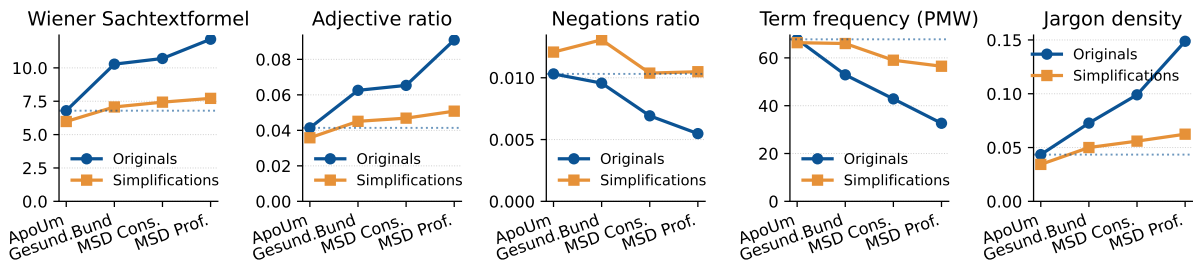

Figure 2: Input complexity bleed-through in simplification outputs for German. On most monotonic metrics, shown here, the simplification reaches roughly Plain Language (*Apotheken Umschau*) level. Note that *Apotheken Umschau* was simplified for completeness, but given it already has a plain language level, this task is redundant.

ing specialized knowledge. This is particularly relevant for German, whose medical lexicon distinguishes sharply between German-rooted terms and Greek/Latin expressions typical of medical reports (e.g. "Blinddarmentzündung" ↔ "Appendizitis"); patients cannot be assumed to recognize these as referring to the same condition. *Apotheken Umschau* bridges this gap with clarifications such as "Der Arzt sagt dazu: Appendizitis." [The doctor calls it: Appendicitis]. *NHS* and *Apotheken Umschau* further reduce terminological precision, yielding shorter, more linear descriptions that preserve the core information while minimizing the need for explanation.

4.3 Simplification Experiment

We probe whether source difficulty leaves a residue in LLM simplifications. Using the cross-source aligned passages, we prompt Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025) to rewrite each text at Plain Language / Einfache Sprache level, then check whether the outputs still track source difficulty on the strictly monotonic readability metrics. A perfect simplifier would erase the ordering; instead, every testable metric except the negation ratio retains it (Figure 2).

However, this bleed-through is weak on average. In addition, the model mostly achieves exactly the metric of the human original Plain Language text. In terms of these metrics, then, simplification works relatively well.

5 Conclusion

We introduce **PlainMedScale**, a topic-aligned medical corpus spanning four comprehensibility levels in German and English. Moving beyond the binary expert–lay contrast, the corpus exposes simplification as a continuum of communicative functions: reference (*MSD Prof.*), explanation (*MSD Consumer*), decision support (*Gesund.Bund*), and access (*Apotheken Umschau* / *NHS*).

Two findings stand out. Most readability metrics established on two registers do not generalize across the full gradient: in German, only the Wiener Sachtextformel, the adjective and negation ratios, and the corpus-based term-frequency and jargon-density measures shift monotonically across all steps. And when asked for Plain Language, a competitive open-weight LLM partially preserves input difficulty, though it roughly reaches the target level. The corpus thus supports stress-testing readability metrics, training, and evaluating systems against multiple targets, and — via the bilingual alignments of *MSD Manuals* and *Gesund.Bund* — studying simplification alongside translation.

Limitations

Authorship and translation direction vary across sources (*Gesund.Bund* authored in German, *MSD Manuals* in English), so cross-lingual comparisons conflate the two. The lowest-difficulty tier is functionally but not editorially parallel: *Apotheken Umschau* follows German Plain Language conventions, while *NHS* uses its own house style. Finally, our alignment pipeline yields no measurable recall figure, since pairs sharing no vocabulary anchor are silently missed. And validation is anchored on Apotheken Umschau pairs only, and we assume the resulting precision transfers to the rest of the corpus.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

Apotheken Umschau. 2020. Apotheken Umschau in Einfacher Sprache. https://www.apotheken-umschau.de/einfache-sprache/.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German Dataset for Joint Summarization and Simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Richard Bamberger and Erich Vanecek. 1984. *Lesen–Verstehen–Lernen–Schreiben*. Jugend und Volk, Wien.

J. Allen Baron, Claudia M. Sánchez-Beato Johnson, Michael A. Schor, Dustin Olley, Lance Nickel, Victor Felix, Susan M. Bello, Carol Greene, Richard Lichenstein, Katharine Bisordi, Rima Koka, Cynthia Bearer, Regina Macatangay, Nischal Ada, Kaitlin Ballenger, Emily Bliss, Lauren Colliver, Grace Dobbins, Harrison Heitzig, Shannan Dixon, Patrick Semesky, Jennifer Garth, Matthew Fairchild, Peter Gaskin, Sarina Zahid, Rachel Castillo, Sarah Edwards, Astrid Widjaja, Yamei Usui, Erin Lynch, Melissa Clarkson, Todd Detwiler, and Lynn M. Schriml. 2026. The DO-KB knowledgebase 2026 update: Expanding programmatic and language access. *Nucleic Acids Research*, 54(D1):D1425–D1436.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. *Preprint*, arXiv:2302.09155.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Bettina M. Bock. 2015. Barrierefreie Kommunikation als Voraussetzung und Mittel für die Partizipation benachteiligter Gruppen. Ein (polito-)linguistischer Blick auf Probleme und Potenziale von «Leichter» und «einfacher Sprache». *Linguistik Online*, 73(4):115–137.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Bundesinstitut für Arzneimittel und Medizinprodukte. 2024. ICD-10-GM Version 2025. Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision, German Modification. https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/_node.html.

Bundesinstitut für Arzneimittel und Medizinprodukte. n.d. DRKS - Deutsches Register Klinischer Studien. https://www.drks.de/search/de.

Council of Europe. 2023. Guide to health literacy. contributing to trust building and equitable access to healthcare. Steering Committee for Human Rights in the fields of Biomedicine and Health (CDBIO). https://rm.coe.int/inf-2022-17-guide-health-literacy/1680a9cb75.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Brigitte Fischer-Wagener, Dietrich Rebholz-Schuhmann, Gabriele Wollnik-Korn, Miriam Albers, Roman Baum, Leyla Jael Castro, and Lukas Geist. 2023. Deutscher MeSH. https://repository.publisso.de/resource/frl%3A6440319.

Gesund.Bund. 2025. Gesund.Bund. https://gesund.bund.de/.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC*, volume 29, pages 34–43.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. Multimsd: A corpus for multilingual medical text simplification from online medical references. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9248–9258, Vienna, Austria. Association for Computational Linguistics.

Inclusion Europe. n.d. Informationen für alle. Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht. Brussels. https://easy-to-read.inclusion-europe.eu/wp-content/uploads/2014/12/DE_Information_for_all.pdf.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-Lite - a New Corpus of Easy German: Compilation, Exploration, Analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, St. Julian's, Malta. Association for Computational Linguistics.

Roman Jakobson. 1960. Closing statement: Linguistics and poetics. In Thomas A. Sebeok, editor, *Style in Language*, pages 350–377. MIT Press, Cambridge, MA.

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

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Association for Computational Linguistics*, pages 6365–6378.

Chen Lyu and Gabriele Pergola. 2024. Society of medical simplifiers. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus. Balancing Comprehensibility and Acceptability*. Frank & Timme, Berlin.

MSD Manuals. 2025. MSD Manuals. https://www.msdmanuals.com/de/.

National Library of Medicine. n.d. PubMed. https://pubmed.ncbi.nlm.nih.gov/.

NHS. 2025. NHS: Health A to Z. https://www.nhs.uk/health-a-to-z/.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Pschyrembel. 2025. Pschyrembel Online. https://www.pschyrembel.de/.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Karen Scholz and Markus Wenzel. 2025. Evaluating Readability Metrics for German Medical Text Simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.

Kristine Sørensen, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, Helmut Brand, and (HLS-EU) Consortium Health Literacy Project European. 2012. Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health*, 12(1):80.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2023. A new aligned simple German corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11393–11412, Toronto, Canada. Association for Computational Linguistics.

U.S. National Library of Medicine. 2025. Medical subject headings (MeSH). https://www.nlm.nih.gov/mesh/.

William Xia, Ishita Unde, Brian Ondov, and Dina Demner-Fushman. 2025. Jebs: A fine-grained biomedical lexical simplification task. *Preprint*, arXiv:2506.12898.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Jargon density

We tokenize each article with SoMaJo (Proisl and Uhrig, 2016) and classify a token as *jargon* when its keyness in medical research texts against a general-language reference corpus exceeds an odds ratio of 20 *and* its relative frequency in that reference corpus is at most 0.001 (the latter filter removes common-but-keyed terms such as *Patient*). Out-of-vocabulary tokens — i.e., tokens absent from the keyness table — are also counted as jargon. The reported metric is *jargon tokens per all tokens*, $N_{\text{jargon}}/N_{\text{tokens}}$. Keyness is precomputed per language: we use DRKS-abstracts (Bundesinstitut für Arzneimittel und Medizinprodukte, n.d.) for German and a sample of PubMed (National Library of Medicine, n.d.) articles for English medical language; and we use Leipzig Newspaper Corpora (Goldhahn et al., 2012) as general-language reference. Because the two corpora differ

in size and register, and because SoMaJo keeps German nominal compounds as single tokens, within-language source comparisons are meaningful but cross-language differences are not.

B Readability scores

Refer to Scholz and Wenzel (2025) for the exact definition of these metrics.

B.1 Abbreviations

FKGL Flesch-Kincaid grade level

CLS score Next-sentence prediction score: how confident BERT is that one sentence naturally follows the previous one (based on BERT's `[CLS]` token)

Term Frequency Mean frequency of content words in general-language corpora. Displayed as **PMW** (per million words) for better readability

Grammar Frequency How many different sentence structures occur in a text, relative to the number of sentences — a text with more repeated structures gets a lower score

Lexical chain A noun that is repeated across a text

NP Noun phrase

Edit distance How much a sentence's structure changes compared to the sentence before it

NP complexity Noun phrase complexity

Fill Mask probability How easily BERT can restore randomly masked words, measured as the true word's normalized rank among all vocabulary candidates. Lower values mean better prediction

Fill-mask probability and average CLS score are computed using BERT[1] for English texts and German BERT[2] for German texts. Average sentence perplexity is computed using GPT-2[3] for English texts and German GPT-2[4] for German texts.

B.2 Full German readability scores

See Table 5 and Figure 3.

B.3 Full English readability scores

See Table 6 and Figure 4.

[1] https://huggingface.co/google-bert/bert-base-uncased
[2] https://huggingface.co/google-bert/bert-base-german-cased
[3] https://huggingface.co/openai-community/gpt2
[4] https://huggingface.co/dbmdz/german-gpt2

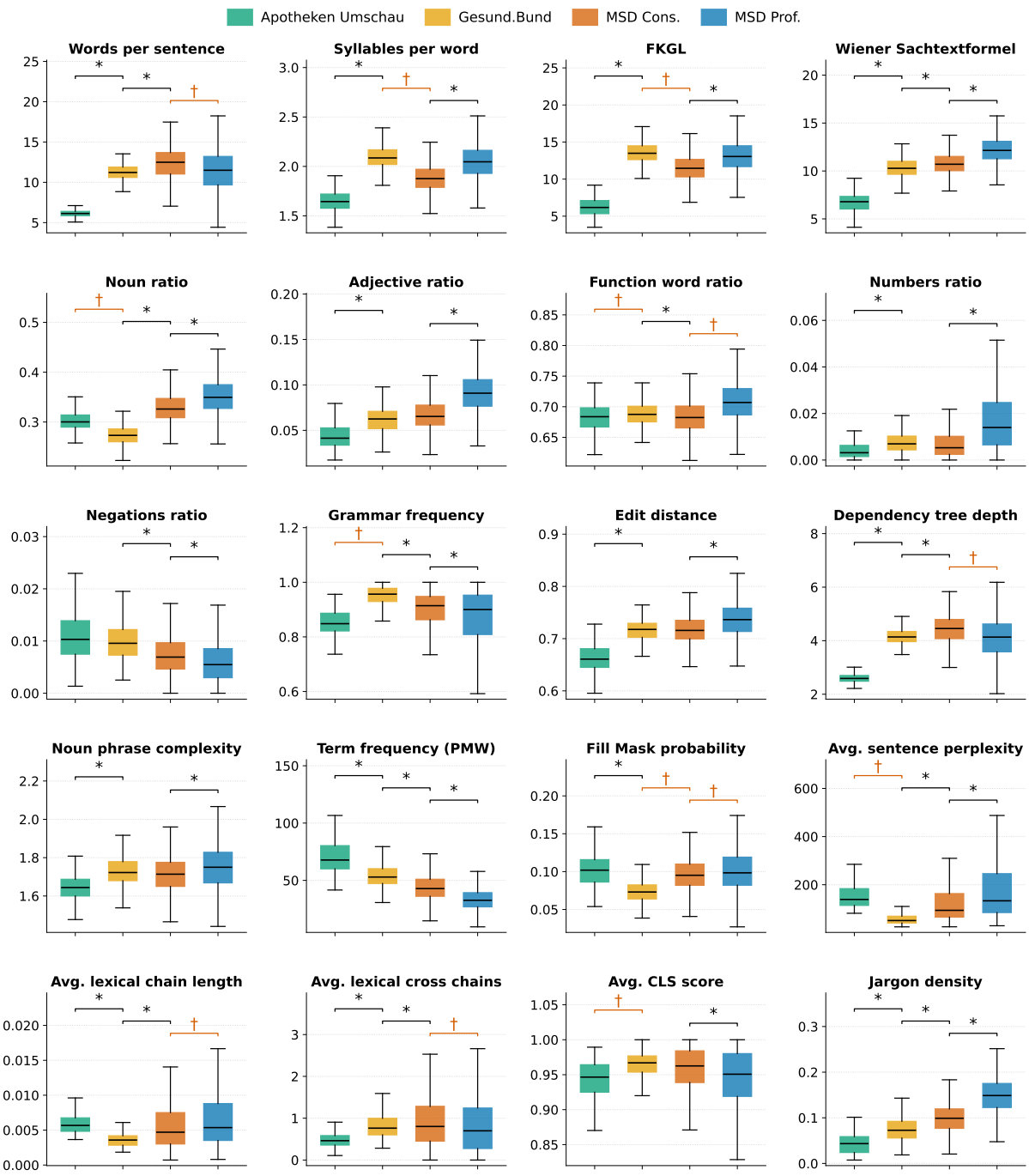

Figure 3: German readability metrics

Metric	Apotheken Umschau		Gesund.Bund		MSD Cons.		MSD Prof.
Words per sentence	6.11 ± 0.45	*	11.23 ± 0.95	*	12.44 ± 2.23	†	11.46 ± 2.90
Syllables per word	1.65 ± 0.11	*	2.10 ± 0.11	†	1.88 ± 0.15	*	2.05 ± 0.20
FKGL	6.27 ± 1.35	*	13.53 ± 1.28	†	11.49 ± 1.79	*	13.11 ± 2.29
Edit distance	0.66 ± 0.03	*	0.72 ± 0.02		0.72 ± 0.03	*	0.73 ± 0.04
Dependency tree depth	2.59 ± 0.17	*	4.15 ± 0.27	*	4.42 ± 0.59	†	4.09 ± 0.82
Noun phrase complexity	1.65 ± 0.08	*	1.73 ± 0.08		1.72 ± 0.10	*	1.75 ± 0.13
Fill Mask probability	0.102 ± 0.022	*	0.074 ± 0.014	†	0.097 ± 0.026	†	0.102 ± 0.033
Avg. sentence perplexity	301 ± 1172	†	68 ± 61	*	288 ± 1234	*	413 ± 2820
Avg. CLS score	0.943 ± 0.027	†	0.963 ± 0.021		0.957 ± 0.039	*	0.942 ± 0.056

Table 5: German readability metrics not in the main article table (i.e., all metrics deemed problematic for German by Scholz and Wenzel (2025)), mean $\pm$ std across articles per source. Markers between adjacent low→high-readability data sources indicate the significant shift in the expected direction (*), or in the opposite direction (†) (one-sided Wilcoxon signed-rank on aligned article pairs, α=0.05, Bonferroni-corrected). Because the test was only done on aligned pairs, full-corpus mean and significance might not agree. Bold/underlined: metrics that are fully/partially significant only in the expected direction.

Metric	NHS		Gesund.Bund		MSD Cons.		MSD Prof.
Words per sentence	10.43 ± 1.41	*	13.22 ± 1.22	*	13.98 ± 2.31	†	13.51 ± 3.13
Syllables per word	1.37 ± 0.09	*	1.59 ± 0.08		1.64 ± 0.10	*	1.81 ± 0.13
FKGL	4.62 ± 1.23	*	8.30 ± 0.93	*	9.16 ± 1.38	*	11.04 ± 1.76
Noun ratio	0.28 ± 0.04	*	0.30 ± 0.03	*	0.35 ± 0.04	*	0.39 ± 0.05
Adjective ratio	0.09 ± 0.02	*	0.11 ± 0.02	†	0.10 ± 0.02	*	0.13 ± 0.03
Function word ratio	0.57 ± 0.03	†	0.59 ± 0.02	†	0.62 ± 0.03	†	0.68 ± 0.04
Numbers ratio	0.01 ± 0.01		0.01 ± 0.01		0.01 ± 0.01	*	0.02 ± 0.02
Negations ratio	0.01 ± 0.01	*	0.01 ± 0.00	*	0.01 ± 0.00	*	0.01 ± 0.00
Grammar frequency	0.94 ± 0.04		0.96 ± 0.03	*	0.91 ± 0.07	*	0.88 ± 0.10
Edit distance	0.74 ± 0.03	†	0.72 ± 0.02		0.72 ± 0.03	*	0.73 ± 0.04
Dependency tree depth	3.78 ± 0.35	*	4.27 ± 0.32	*	4.51 ± 0.55	†	4.34 ± 0.74
NP complexity	1.67 ± 0.13	*	1.91 ± 0.11	†	1.84 ± 0.13	*	1.94 ± 0.16
Term frequency (PMW)	141.2 ± 28.9	*	116.7 ± 27.0		108.2 ± 30.2	*	64.0 ± 21.5
Fill Mask probability	0.005 ± 0.006		0.006 ± 0.005	*	0.002 ± 0.004	†	0.003 ± 0.004
Avg. sentence perplexity	123 ± 151		89 ± 64	†	93 ± 226	*	141 ± 501
Avg. lexical chain length	0.01 ± 0.00	*	0.00 ± 0.00		0.01 ± 0.00	†	0.01 ± 0.01
Avg. lexical cross chains	0.58 ± 0.27	*	1.09 ± 0.41	*	1.24 ± 0.81	*	1.46 ± 1.12
Avg. CLS score	0.958 ± 0.041		0.971 ± 0.015	†	0.980 ± 0.032		0.980 ± 0.038
Jargon density	0.040 ± 0.029	*	0.063 ± 0.032	*	0.094 ± 0.042	*	0.167 ± 0.063

Table 6: English readability metrics, mean $\pm$ std across articles per source. Markers between adjacent low→high-readability data sources indicate the significant shift in the expected direction (*), or in the opposite direction (†) (one-sided Wilcoxon signed-rank on aligned article pairs, α=0.05, Bonferroni-corrected). Because the test was only done on aligned pairs, full-corpus mean and significance might not agree. Bold/underlined: metrics that are fully/partially significant only in the expected direction.

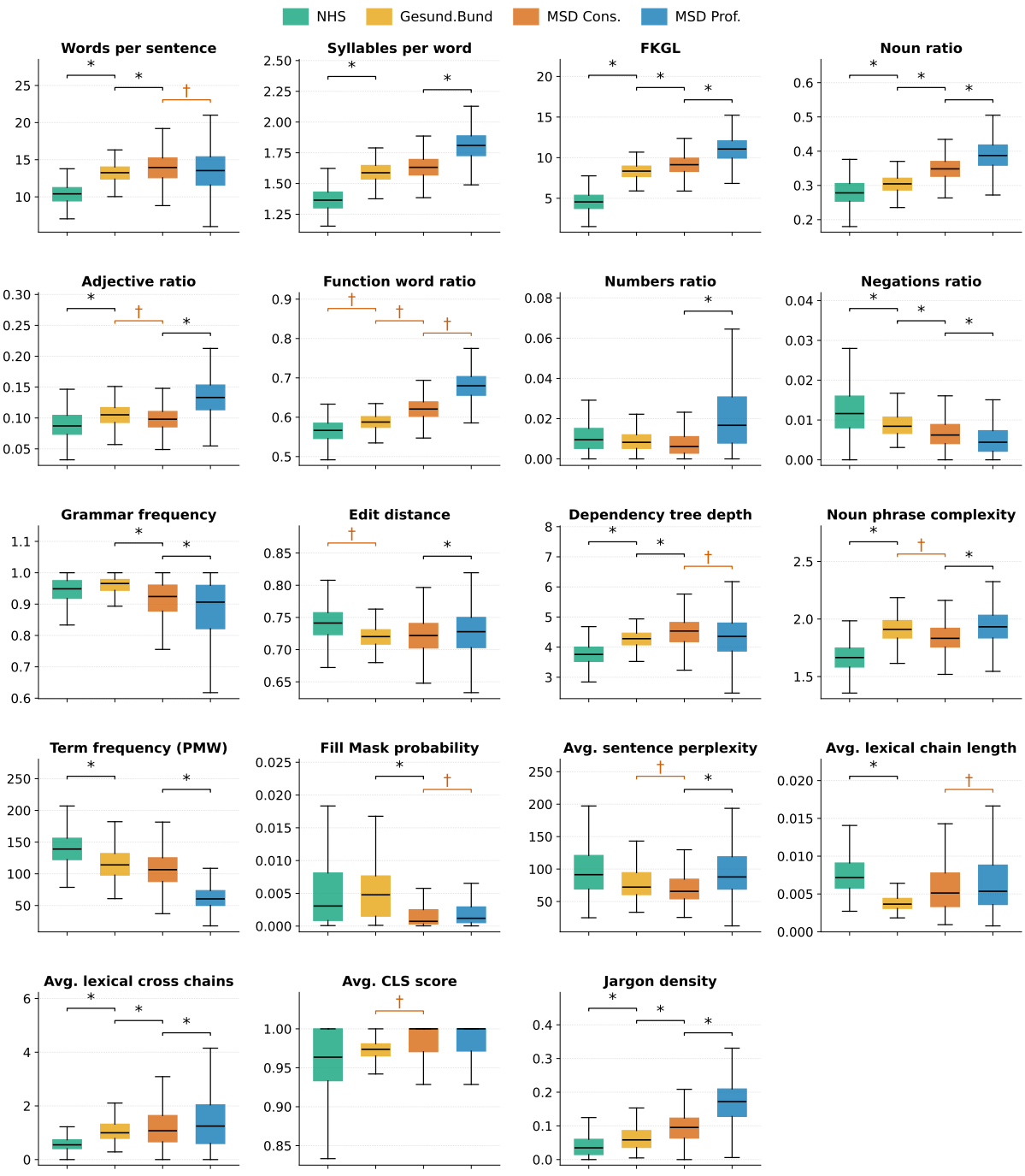

Figure 4: English readability metrics