

Evaluation of Multilingual Text Simplification for the Mental Health Domain: Exploring Small Language Models

Olga Pelloni¹, Sandra Anna Just^{2,3}, Lars Ailo Bongo¹

¹Department of Computer Science, UiT The Arctic University of Norway,

²Department of Clinical Medicine, UiT The Arctic University of Norway,

³Department of Psychiatry and Neurosciences, Charité—Universitätsmedizin, Berlin, Germany

{olga.pelloni, sandra.a.just, lars.ailo.bongo}@uit.no

Abstract

Individuals with particular mental health disorders may find it difficult to learn about their own condition. Therefore, efforts have been made to create materials that explain complex medical information in simpler words, which are also beneficial for caregivers and others. However, text simplification is commonly done in English and only sporadically in other languages. In this study, we explore potential ways for language-agnostic medical text simplification for the mental health domain. Our approach is to simplify the ICD-11 articles on primary psychotic disorders in English, German and French, using small LMs and various metrics for evaluating different aspects of the texts: lexical complexity and readability. Our results show that acceptable texts were produced only in English, and that a joint analysis of Measure of Textual Lexical Diversity (MTLD) and Flesch Reading Ease (FRE) provides the most insight, capturing both the best outcomes and signaling different types of issue. The study is preliminary and requires further investigation.

1 Introduction

As with any professional or technical domain, medical texts tend to be difficult to understand for non-specialists. Clinical NLP explores methods for different transformations of medical texts (Hasan and Farri, 2019), including *paraphrase generation* and *text simplification*, where the latter is a more complex task even in English (Basu et al., 2023).

Simplification of medical texts in languages other than English remains understudied. Some research on French and Romanian focuses primarily on paraphrase generation and shows good results when using language- and domain-specific models (Buhnla and Todirascu, 2024; Buhnla et al., 2024; Buhnla, 2025), whereas research on text simplification in Polish and some Indian languages, Arabic and Swahili reports that general purpose LLMs, such as GPT-4, provide better results than smaller

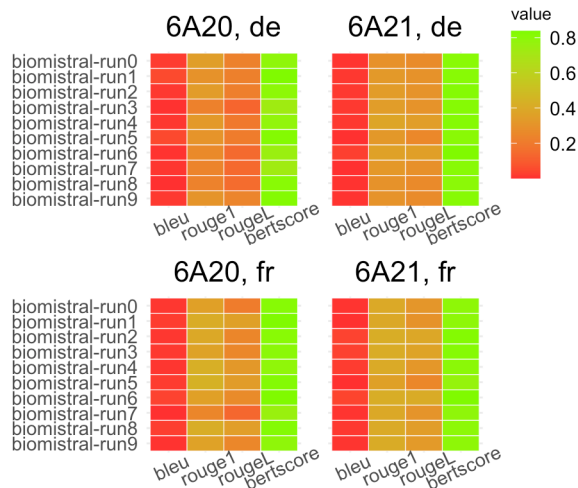


Figure 1: Typically used similarity metrics BLEU, ROUGE and BERTScore do not reflect fine-grained differences in simplified texts when comparing different languages, model runs and ICD-11 articles, although the resulting texts vary greatly in terms of fluency and simplicity (our results).

fine-tuned models (Grzybowski et al., 2025; Gan-gavarapu, 2024).

In contrast to text simplification in primary medicine, the mental health domain requires further attention to potential cognitive impairments in some mental health disorders, neutral and non-stigmatizing tone, simplifying both words and sentences without making text unnatural. These nuances can be challenging to account for in multilingual settings, even when translating medical texts manually (Natoli, 2026).

Given the ambitious nature of the outlined task, we conduct a preliminary study to investigate evaluation methods capable of detecting various issues in a multilingual setting. In this study, we explore small models, intentionally including some negative results (poor-quality texts), to focus on identifying key metrics that highlight text properties of the best and worst simplified texts across 3 languages.

Small models are chosen for greater reproducibility in a metric exploration study.

2 Related Work

The need for reading materials in simplified language has long been recognized, especially in healthcare contexts. Different lay persons are considered beneficiaries of so-called Easy Language, and various practices for it have been established across countries for manual adaptation of texts (Nomura et al., 2010; Lindholm and Vanhatalo, 2021).

Despite these efforts, automatic methods for clinical text simplification remain heavily focused on English, with only limited work on other European languages (Buhnila and Todirascu, 2024; Buhnila et al., 2024; Buhnila, 2025; Grzybowski et al., 2025; Scholz and Wenzel, 2025) and even less beyond Europe (Gangavarapu, 2024).

Automatic simplification of medical texts raises ethical considerations, such as the risk of meaning distortion. (Gooding, 2022) This is even more important when targeting texts in the psychiatric domain (Aich et al., 2025) and audiences with mental health disorders (Gralton et al., 2010), given the long history of stigmatization (Rössler, 2016). Evaluation of the resulting texts typically relies on metrics coming from machine translation (MT) that measure similarity between the original and the simplified texts on the level of word tokens (Papineni et al., 2002; Lin, 2004) and on semantic level (Zhang et al., 2019), which provide only limited insight on the quality of the simplified texts (Figure 1). The Flesch Reading Ease metric (Flesch, 1948) is widely used to assess readability, while the other readability metrics and other properties of text, such as lexical complexity, are typically overlooked.

Our preliminary study addresses these gaps by analyzing 3 lexical complexity and 9 readability metrics to identify more informative and potentially language-agnostic evaluation methods.

3 Data & Methods

We extract five articles from ICD-11 (Harrison et al., 2021) using ICD API¹, containing definitions of primary psychotic disorders: 6A20–6A24 including all the sub-definitions (6A20.0, 6A20.1, etc.). The concatenated definitions serve as input to the following small language models, fine-tuned for text simplification and/or for multilingual use:

¹<https://icd.who.int/icdapi>

- **LingConv** (Elgaar and Amiri, 2025): a text conversion tool² based on a T5-Base (Raffel et al., 2020) model intended for general paraphrase generation in English;
- **PRISM** (Thompson and Post, 2020): a transformer-based MT model adapted for general paraphrase generation, trained on 39 languages;
- **Ascle** (Yang et al., 2024): a toolkit for medical text generation, incl. a text simplification function. The toolkit includes 3 models (BigBird-Pegasus Large (Zhang et al., 2020); BART Large (Lewis et al., 2020) and BioBART v2 Base (Yuan et al., 2022)), each in 2 versions fine-tuned on different medical data, primarily in English.
- **BioMistral 7B** (Labrak et al., 2024): a Mistral 7B Instruct model (Jiang et al., 2023) fine-tuned for biomedical domain in 9 languages. We use the BioMistral SLERP model³ with 4-bit GPTQ quantization (Frantar et al., 2022).

All models except BioMistral were given the ICD-11 articles as input, while for BioMistral we provided a simple prompt: "Rewrite this medical definition for a layman person. Keep it as close to the original meaning as possible, just use simpler words and syntactic constructions."; we manually translated it into German and French.

We test all the models on English texts first and calculate the following evaluation metrics:

- **Similarity:** BLEU, ROUGE 1 and ROUGE L, BERTScore. These are standard metrics in MT for assessing the preservation of the n-gram combinations and the underlying meaning;
- **Lexical complexity:** TTR (type-token ratio), MATTR (moving-average type-token ratio) (Covington and McFall, 2010), MTLT (measure of textual lexical diversity) (McCarthy and Jarvis, 2010). All of these measures are based on the ratio of unique words (types) and all word instances (tokens) in a given text;
- **Readability:** Flesch Reading Ease (FRE), Flesch-Kincaid grade level (FKGL), Automated Readability Index (ARI) (Kincaid et al.,

²<https://mohdelgaar-lingconv.hf.space/>

³<https://huggingface.co/LoneStriker/BioMistral-7B-SLERP-GPTQ>

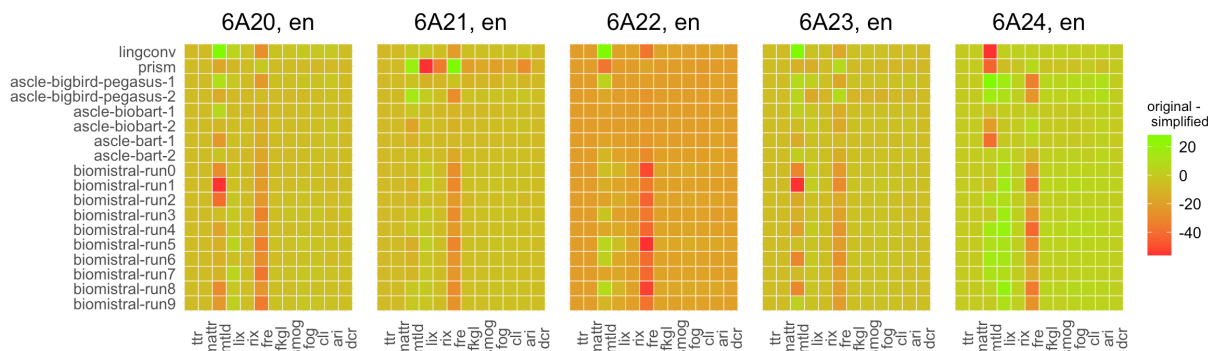


Figure 2: Difference in complexity and readability metrics across 9 models and 5 ICD-11 articles in English.

1975), SMOG index (SMOG) (Mc Laughlin, 1969), Gunning Fog index (Fog) (Gunning, 1952), Coleman–Liau Index (CLI) (Coleman and Liau, 1975), Dale-Chall Readability (DCR) (Chall and Dale, 1995), Läsbarhetsindex (LIX) (Björnsson, 1968) and Readability Index (RIX) (Anderson, 1983). These metrics capture length of words, sentences; some of them also check a set of familiar words (e.g. Dale-Chall).

Given the results on English, we manually evaluate the simplified texts from each model and identify the most aligned metrics. The reduced set of metrics is further used to evaluate the results on German and French, taking into account that some of the readability metrics cannot be easily transferred to languages other than English.⁴ In addition, three human annotators evaluated BioMistral results in English.

We use textstat (Bansal and Aggarwal, 2026) and LexicalRichness (Shen, 2022) Python libraries to calculate all the readability and lexical complexity metrics. In particular, we observe the *differences* in these metrics between the original and simplified texts, so that rather than relying on absolute metric values, we assess the *reduction in complexity*, which allows us to do better comparisons cross-linguistically. Our code is open sourced and available at: <https://github.com/uit-hdl/multilingual-text-simplification-icd-11>

4 Results & Discussion

Figure 2 shows the differences in complexity and readability metrics measured on English texts. The general paraphrasing models LingConv and PRISM produce the worst texts, but in different

⁴The library textstat provides adaptations of FRE and Fog to several languages: <https://pypi.org/project/textstat/>.

Rank	#1	#2	#3
Best (1)	4, 0	8, 7	0
Good (2)	9, 8	9, 6	1, 2
Medium (3)	7, 6	2, 3	3, 7, 8, 9
Poor (4)	5, 1	4, 5	4, 6
Worst (5)	3, 2	1, 0	5

Table 1: Human experts’ rankings of the texts in English, BioMistral model’s runs 0–9, grouped into categories from 1 to 5. We mark runs with ≥ 2 annotator agreement and no opposing third (Cohen’s $\kappa \approx 0$).

ways: LingConv tends to produce just one short and simple sentence (e.g. "Schizophrenia is known as a disease."), while PRISM tends to keep sentences very close to the original definitions ("Schizophrenia can be characterized by disorders").

Thus, the first two rows stand for extreme cases, where LingConv is too simple and PRISM is too complex. Ascle models produced texts of mediocre quality, somewhat close to PRISM but not as complex (e.g. "Schizophrenia is a mental illness that is characterized by changes").⁵ BioMistral model produced the most natural simplified texts when given 6A20 article as input compared to the other models (e.g. "Schizophrenia is a mental illness that affects how you think, feel, and behave").

Observing the differences in metrics, we notice that MTLT (3rd column in Figure 2) and FRE (6th column) are the most prominent features⁶. The best simplified texts of the 6A20 article in English produced by BioMistral are characterized by increase in MTLT (in red) with a *simultaneous* increase in FRE (in red). The only results with substantial increase of these two metrics were produced by BioMistral when processing 6A20 and 6A23

⁵See original texts in Appendix, Table 2.

⁶Note that some of the metrics cannot show big difference in values by nature, e.g. based on grade level; we observed each of these metrics separately and did not find any consistent patterns with respect to the quality of the models’ output.

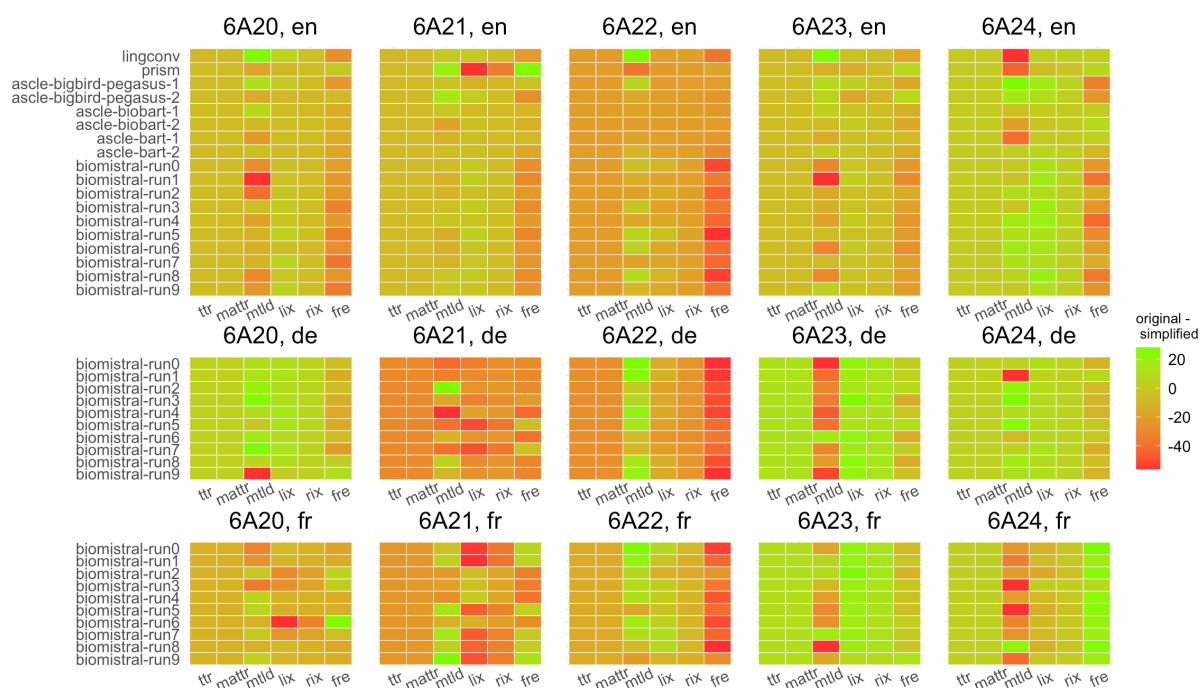


Figure 3: Difference in complexity and readability metrics between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

articles.

The other models mostly produce texts which either show a decrease in MTLT (in green) and increase in FRE (in red) or vice versa, or no change compared to the original texts.

We did an additional check by three human experts to evaluate the texts produced by BioMistral in English for 6A20 definition to see if the fluctuations in MTLT and FRE align with human judgments (Table 1). While the human experts reported that generally the texts look similar to each other, they could identify some best and worst results within the BioMistral runs: at least two of the experts mark runs 8, 9 as the best and 1, 5 as the worst. Best results have a moderate increase in MTLT (in the range of 27–37), unlike the worst texts, where the increase in MTLT is very high (82 in run 1) or very low (8 in the run 5) (see examples in Appendix, Table 3).

Figure 3 shows the differences in complexity metrics across English, German and French. We only present results of the BioMistral model, as the other multilingual model PRISM did not produce any paraphrases neither in German nor in French; in fact, some English outputs contained mixed expressions with French or Spanish (e.g. "caractérisated"). BioMistral in German and French struggles more than in English: texts mostly remain complex, contain grammatical mistakes and are not

smooth. This is reflected by increase in MTLT (in red) and decrease in FRE (in green) and vice versa (see examples in Appendix, Table 4, 5).

Some poor-quality results might be due to the inherent properties of the input text, e.g. 6A22 is a short article, thus, not much transformation is happening in all three languages: FRE increases as expected, but MTLT consistently decreases. This is also noticeable in differences of the number of word tokens and types (see Appendix, Figure 4).

5 Conclusion

In this study, we explored 3 lexical complexity and 9 readability metrics to evaluate multilingual text simplification in the mental health domain. We find that a combination of differences in MTLT and FRE helps evaluating simplified texts: increase of FRE means that words become shorter, and increase in MTLT makes the lexical diversity higher, while decrease in MTLT results in repetitive words. Increase in MTLT and decrease in FRE signify the opposite: words become longer and there is high lexical diversity, making texts too dense. In addition, small LMs worked well only on English, while German and French resulted in poor-quality texts. We suggest exploring these metrics for medical text simplification further with LLMs and using them as a pre-check for human evaluation.

Limitations

This is a preliminary study, and we acknowledge that stronger claims could be made if we had extended our analysis to more models, languages and text samples. We report mostly negative results in terms of the quality of the simplified texts; however, we intentionally limit this study to small models to allow for better control and exploration of the evaluation metrics. We recognize that higher-quality texts and more effective use of the metrics would likely be achieved when using large language models.

References

- Ankit Aich, Tingting Liu, Salvatore Giorgi, Kelsey Isman, Ruhshana Bobojonova, Lyle Ungar, and Brenda Curtis. 2025. Language models in digital psychiatry: challenges with simplification of healthcare materials. *NPP—Digital Psychiatry and Neuroscience*, 3(1):10.
- Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.
- Shivam Bansal and Chaitanya Aggarwal. 2026. Textstat: Python readability statistics library. <https://github.com/textstat/textstat>.
- Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14093–14101.
- Carl-Hugo Björnsson. 1968. *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.
- Ioana Buhnila. 2025. Explain this medical term in my language: A case study of small language models for medical paraphrase generation. In *3rd UniDive Workshop*, pages 1–4, Budapest, Hungary.
- Ioana Buhnila, Aman Sinha, and Mathieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.
- Ioana Buhnila and Amalia Todirascu. 2024. **LARGEMED: A resource for identifying and generating paraphrases for French medical terms**. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 141–151, Torino, Italia. ELRA and ICCL.
- Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Michael A Covington and Joe D McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, 17(2):94–100.
- Mohamed Elgaar and Hadi Amiri. 2025. **LingConv: An interactive toolkit for controlled paraphrase generation with linguistic attribute control**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–51, Suzhou, China. Association for Computational Linguistics.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.
- Agasthya Gangavarapu. 2024. Introducing L2M3, a multilingual medical large language model to advance health equity in low-resource regions. *arXiv preprint arXiv:2404.08705*.
- Sian Gooding. 2022. **On the ethical considerations of text simplification**. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.
- Ernest Galton, Marilyn Sher, and Coro Drew Lopez. 2010. Information and readability issues for psychiatric patients: e-learning for users. *The Psychiatrist*, 34(9):376–380.
- Łukasz Grzybowski, Jakub Pokrywka, Michał Ciesiółka, Jeremi Ignacy Kaczmarek, and Marek Kubis. 2025. **Polish-English medical knowledge transfer: A new benchmark and results**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9042–9063, Suzhou, China. Association for Computational Linguistics.
- R. Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.
- James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. ICD-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21(Suppl 6):206.
- Sadid A Hasan and Oladimeji Farri. 2019. Clinical natural language processing with deep learning. In *Data Science for Healthcare: Methodologies and Applications*, pages 147–171. Springer.

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. *Mistral 7B*. Preprint, arXiv:2310.06825.
- 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.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. *BioMistral: A collection of open-source pretrained large language models for medical domains*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Camilla Lindholm and Ulla Vanhatalo. 2021. *Handbook of easy languages in Europe*. Frank & Timme.
- G Harry Mc Laughlin. 1969. SMOG grading – a new readability formula. *Journal of reading*, 12(8):639–646.
- Philip M McCarthy and Scott Jarvis. 2010. MTL D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- Adam P Natoli. 2026. Leveraging artificial intelligence to linguistically compare test translations: A methodological introduction and demonstration. *Assessment*, 33(2):163–177.
- Misako Nomura, Gyda Skat Nielsen, and Bror Tronbacke. 2010. Guidelines for easy-to-read materials. *IFLA professional reports*, (120):0_1.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Wulf R ssler. 2016. The stigma of mental disorders: A millennia-long history of social exclusion and prejudices. *The EMBO Reports*, 17(9):1250–1253.
- 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.
- Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness. <https://github.com/LSYS/lexicalrichness>. DOI: 10.5281/zenodo.6607007.
- Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*, Online. Association for Computational Linguistics.
- Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha Dave, Tiarnan Keenan, and 1 others. 2024. Ascle—a python natural language processing toolkit for medical text generation: development and evaluation study. *Journal of medical Internet research*, 26:e60601.
- Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaying Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

A Appendix

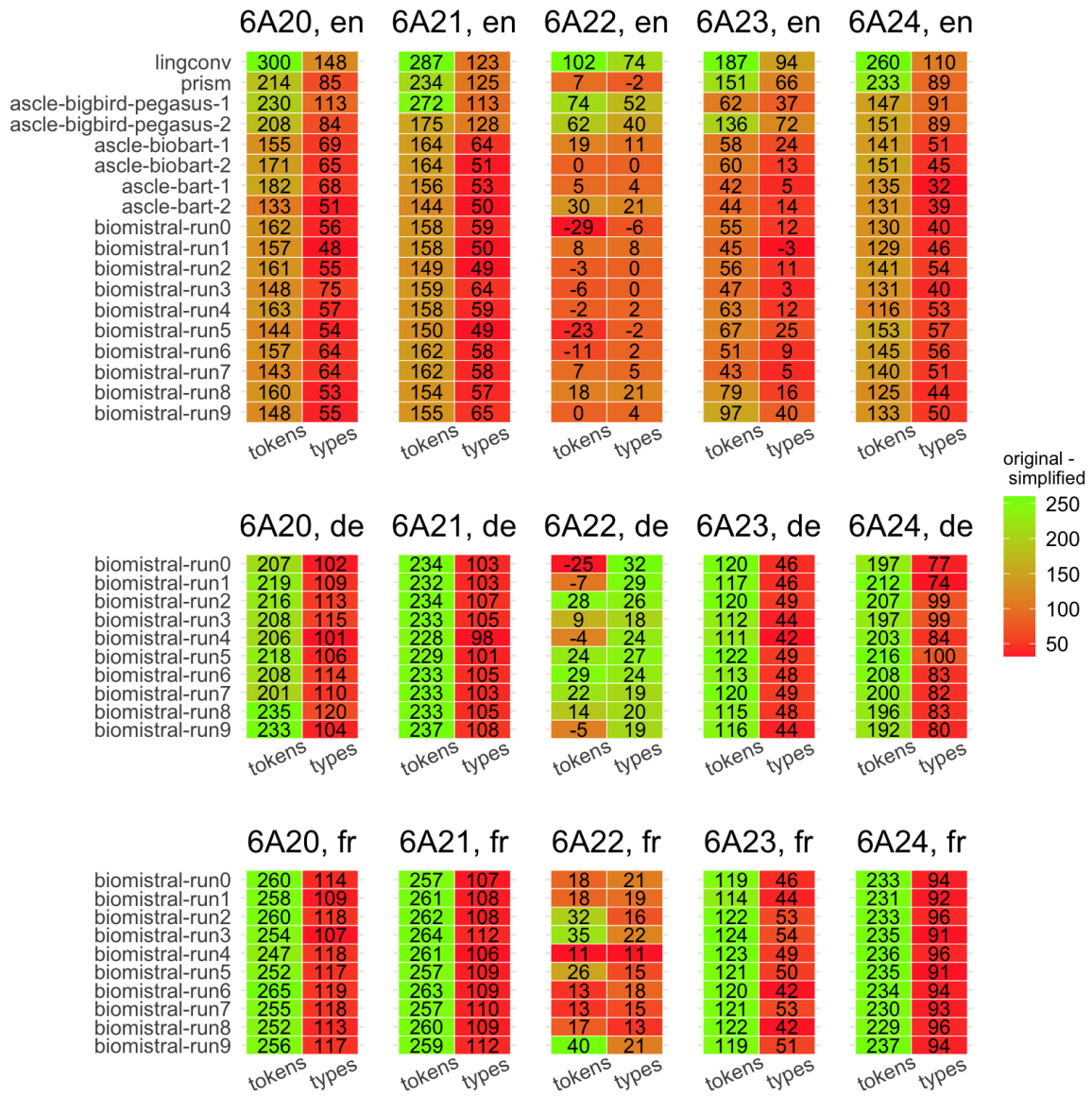


Figure 4: Difference in number of word tokens and word types (unique words) between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

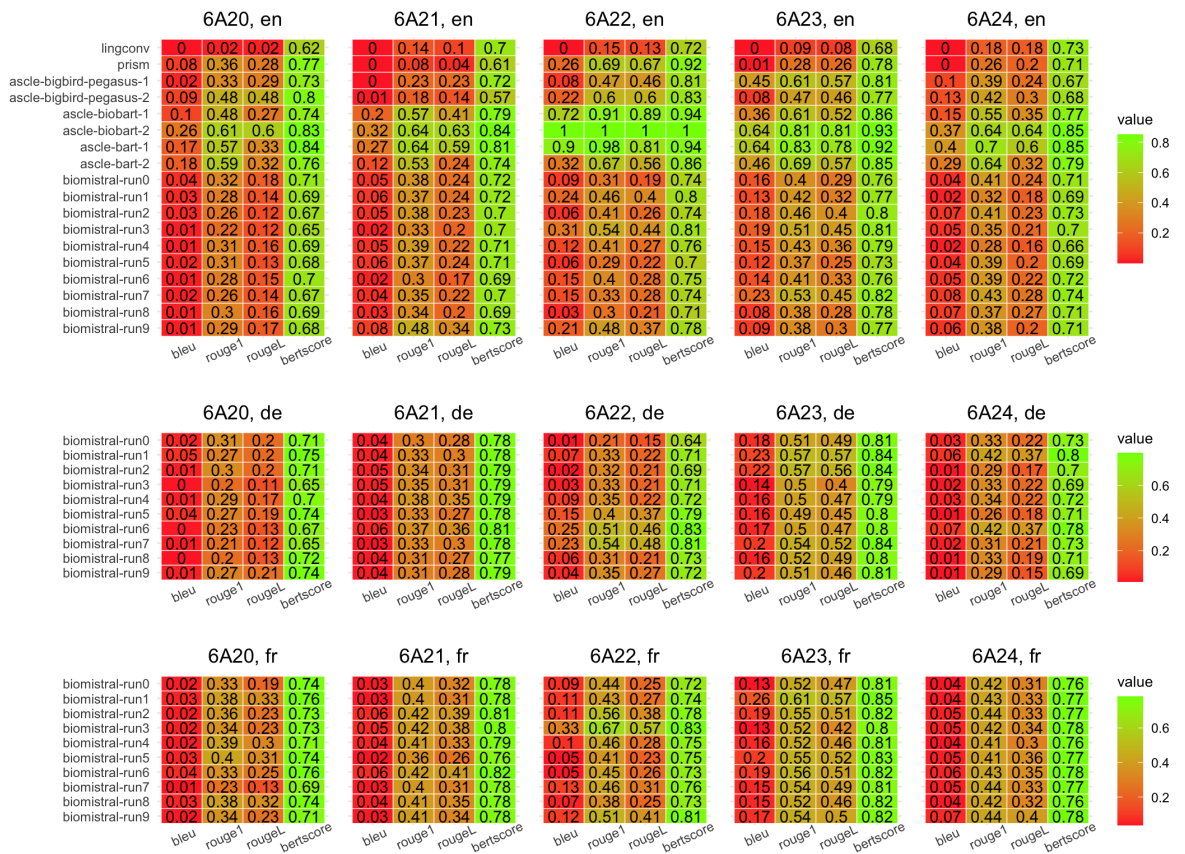


Figure 5: Similarity metrics between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

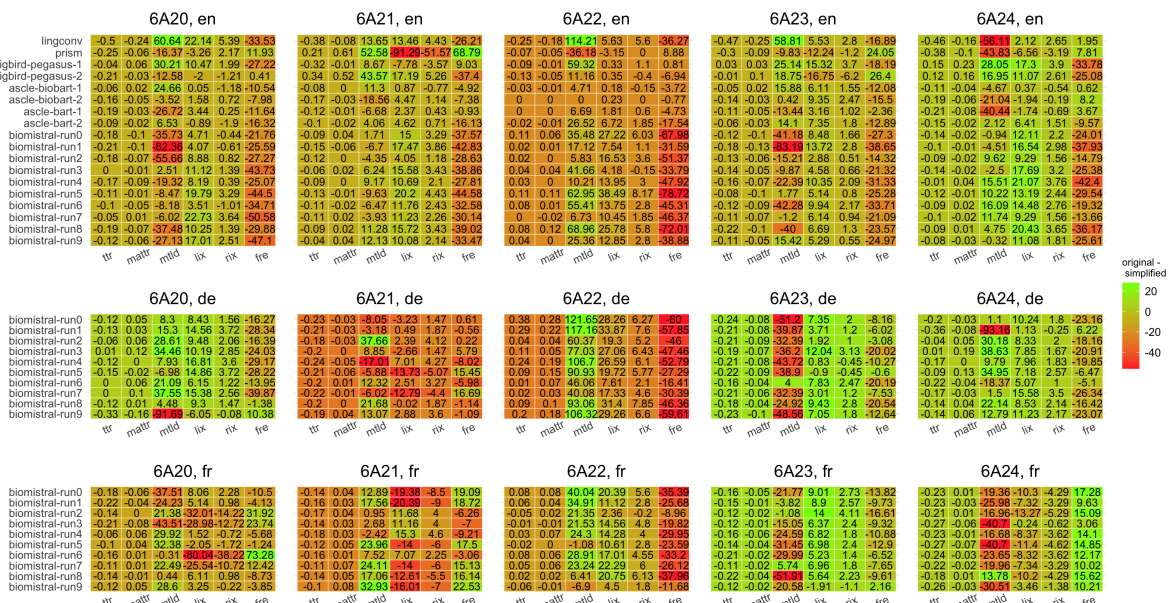


Figure 6: Difference in complexity and readability metrics between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

Original text (en)	Original text (de)	Original text (fr)
<p>Schizophrenia is characterised by disturbances in multiple mental modalities, including thinking (e.g., delusions, disorganisation in the form of thought), perception (e.g., hallucinations), self-experience (e.g., the experience that one's feelings, impulses, thoughts, or behaviour are under the control of an external force), cognition (e.g., impaired attention, verbal memory, and social cognition), volition (e.g., loss of motivation), affect (e.g., blunted emotional expression), and behaviour (e.g., behaviour that appears bizarre or purposeless, unpredictable or inappropriate emotional responses that interfere with the organisation of behaviour). Psychomotor disturbances, including catatonia, may be present. Persistent delusions, persistent hallucinations, thought disorder, and experiences of influence, passivity, or control are considered core symptoms. Symptoms must have persisted for at least one month in order for a diagnosis of schizophrenia to be assigned. The symptoms are not a manifestation of another health condition (e.g., a brain tumour) and are not due to the effect of a substance or medication on the central nervous system (e.g., corticosteroids), including withdrawal (e.g., alcohol withdrawal). Schizophrenia, first episode should be used to identify individuals experiencing symptoms that meet the diagnostic requirements for Schizophrenia (including duration) but who have never before experienced an episode during which diagnostic requirements for Schizophrenia were met. Schizophrenia, multiple episodes should be used to identify individuals experiencing symptoms that meet the diagnostic requirements for Schizophrenia and who have also previously experienced episodes during which diagnostic requirements were met, with substantial remission of symptoms between episodes.</p>	<p>Schizophrenie ist durch Störungen in mehreren mentalen Modalitäten gekennzeichnet, einschließlich des Denkens (z. B. Wahnsymptome, Desorganisation in der Form des Denkens), der Wahrnehmung (z. B. Halluzinationen), des Selbsterlebens (z. B. das Erleben, dass die eigenen Gefühle, Impulse, Gedanken oder das Verhalten unter der Kontrolle einer externen Kraft stehen), der Kognition (z. B., Beeinträchtigung der Aufmerksamkeit, des verbalen Gedächtnisses und der sozialen Kognition), des Willens (z. B. Motivationsverlust), des Affekts (z. B. reduzierter Gefühlsausdruck) und des Verhaltens (z. B. Verhalten, das bizarr oder zwecklos erscheint, unvorhersehbare oder unangemessene emotionale Reaktionen, die die Organisation des Verhaltens stören). Psychomotorische Störungen, einschließlich Katatonie, können vorhanden sein. Anhaltende Wahnvorstellungen, anhaltende Halluzinationen, Denkstörungen und Beeinflussungserleben gelten als Kernsymptome. Die Symptome müssen mindestens einen Monat lang bestanden haben, damit die Diagnose Schizophrenie gestellt werden kann. Die Symptome sind nicht Ausdruck eines anderen Gesundheitszustands (z. B. eines Hirntumors) und sind nicht auf die Wirkung einer Substanz oder eines Medikaments auf das zentrale Nervensystem (z. B. Kortikosteroide) zurückzuführen, auch nicht auf einen Entzug (z. B. Alkoholentzug). Schizophrenie, erste Episode sollte verwendet werden, um Personen zu identifizieren, deren Symptome die diagnostischen Anforderungen für Schizophrenie (einschließlich Dauer) erfüllen, die aber noch nie eine Episode erlebt haben, in der die diagnostischen Anforderungen für Schizophrenie erfüllt wurden. Schizophrenie, mehrfache Episoden sollte verwendet werden, um Personen zu identifizieren, die Symptome aufweisen, die die diagnostischen Anforderungen für Schizophrenie erfüllen, und die auch schon früher Episoden erlebt haben, in denen die diagnostischen Anforderungen erfüllt waren, wobei die Symptome zwischen den Episoden deutlich zurückgegangen sind.</p>	<p>La schizophrénie se caractérise par des perturbations de multiples modalités mentales, notamment la réflexion (p. ex. idées délirantes, désorganisation dans la forme de la pensée), la perception (p. ex. hallucinations), l'expérience de soi (p. ex. l'impression que ses propres sentiments, impulsions, pensées ou comportements sont sous le contrôle d'une force externe), la cognition (p. ex. trouble de l'attention, de la mémoire verbale et de la cognition sociale), la volition (p. ex. perte de motivation), l'affect (p. ex. expression émotionnelle éteinte) et le comportement (p. ex. comportement qui semble bizarre ou inutile, réactions émotionnelles imprévisibles ou inappropriées qui perturbent l'organisation du comportement). Des troubles psychomoteurs, y compris une catatonie, peuvent être présents. Les idées délirantes persistantes, hallucinations persistantes, troubles de la pensée et expériences d'influence, de passivité ou de contrôle sont considérés comme des symptômes fondamentaux. Les symptômes doivent avoir persisté pendant au moins un mois pour qu'un diagnostic de schizophrénie soit posé. Les symptômes ne sont pas une manifestation d'une autre affection médicale (p. ex. une tumeur cérébrale) et ne sont pas dus à l'effet d'une substance ou d'un médicament sur le système nerveux central (p. ex. corticostéroïdes), y compris à un sevrage (p. ex. sevrage alcoolique). La schizophrénie, premier épisode doit être utilisé pour identifier les individus qui présentent des symptômes satisfaisant les critères de diagnostic de la schizophrénie (y compris la durée) mais qui n'ont jamais présenté d'épisode précédent durant lequel les critères de diagnostic de la schizophrénie étaient satisfaits. Schizophrénie, épisodes multiples doit être utilisé pour identifier les personnes présentant des symptômes qui répondent aux critères diagnostiques de la schizophrénie et qui ont également déjà connu des épisodes au cours desquels les critères diagnostiques ont été remplis, avec une rémission substantielle des symptômes entre les épisodes.</p>

Original text (en)	Original text (de)	Original text (fr)
Some attenuated symptoms may remain during periods of remission, and remissions may have occurred in response to medication or other treatment. Symptoms fulfilling all definitional requirements of Schizophrenia have been present for almost all of the illness course over a period of at least one year, with periods of subthreshold symptoms being very brief relative to the overall course.	Einige abgeschwächte Symptome können während der Remissionsphasen bestehen bleiben, und Remissionen können als Reaktion auf Medikamente oder andere Behandlungen eingetreten sein. Die Symptome, die alle definitorischen Anforderungen der Schizophrenie erfüllen, waren fast während des gesamten Krankheitsverlaufs über einen Zeitraum von mindestens einem Jahr vorhanden, wobei die Perioden mit unter-schweligen Symptomen im Vergleich zum Gesamtverlauf sehr kurz waren.	Certains symptômes atténués peuvent persister pendant les périodes de rémission, et des rémissions peuvent se produire en réponse à des médicaments ou à un autre traitement. Des symptômes satisfaisant tous les critères de définition de la schizophrénie ont été présents pendant la quasi-totalité de l'évolution de la maladie sur une période d'au moins un an, avec des périodes de symptômes en-dessous du seuil très brèves par rapport à l'évolution globale.

Table 2: Original ICD-11 6A20 description of Schizophrenia (incl. 6A20.0, 6A20.1, 6A20.2) in English, German and French.

Simplified text (en, BioMistral, run 8)	Simplified text (en, BioMistral, run 5)
Schizophrenia is a mental illness that affects many different parts of the mind. It can cause problems with thinking, feeling, and behaving. People with schizophrenia might have strange thoughts or beliefs (delusions), see or hear things that aren't really there (hallucinations), feel like their thoughts or actions are being controlled by someone else, or have trouble remembering things or paying attention. They might also feel less motivated, have trouble controlling their emotions, or behave in ways that seem strange or out of place. Sometimes, people with schizophrenia might become stuck in one position for a long time (catatonia). To be diagnosed with schizophrenia, a person must have had symptoms for at least a month that are not caused by another medical condition or by drugs or alcohol. Schizophrenia is usually divided into three categories: first episode, multiple episodes with remission, and persistent. First episode schizophrenia.	Schizophrenia is a mental illness that affects the way you think, feel, act, and sense things. It can cause problems with your thoughts, like having false beliefs or trouble organizing your thoughts. It can also affect your senses, like hearing or seeing things that aren't really there. You may feel like your emotions are out of control or that you're not in control of your own actions. You might have trouble remembering things or paying attention. You may lose motivation or interest in things. People with schizophrenia may behave in strange or unusual ways, like being overly emotional or having trouble controlling their body movements. They may also have trouble understanding social situations or other people's feelings. Psychomotor disturbances , like being stuck in one position for a long time, may be present. To be diagnosed with schizophrenia, you must have symptoms that have lasted for at least a month and that are not caused by another medical condition or

Table 3: Examples of simplified texts in English, 6A20, run 8 (with high increase in MTL D by 37.48 and increase in FRE by 29.88, ranked as best by 2 human annotators) and run 5 (with lower increase in MTL D by 8.47 and increase in FRE by 44.5, rated as worst by 3 human annotators). The second text shows more repetitive structures (highlighted) reflected by smaller difference in MTL D; the text also has a complex medical term "psychomotor disturbances". The truncated sentences at the end are due to limit in tokens number; they were not manually edited.

Simplified text (de, BioMistral, run 9)	Simplified text (de, BioMistral, run 7)
<p>Schizophrenie ist eine Krankheit, bei der Menschen ihre Denkweise, Wahrnehmung, Gefühle und Verhalten stark beeinträchtigt werden. Diese Beeinträchtigungen können in verschiedenen Formen auftreten, wie Wahnvorstellungen, Halluzinationen, das Erleben, dass ihre eigenen Gefühle, Impulse, Gedanken oder Verhalten unter der Kontrolle einer externen Kraft stehen, Beeinträchtigung ihrer Kognition (z. B. Aufmerksamkeit, Gedächtnis, soziale Fähigkeiten), Motivationsverlust, reduzierter Gefühlsausdruck und bizarre oder zweckloses Verhalten. Psychomotorische Störungen, wie Katatonie, können ebenfalls vorhanden sein. Für die Diagnose Schizophren</p>	<p>Schizophrenie ist eine Krankheit, bei der man sich nicht sicher ist, was man denkt, was man sieht oder hört, wie man sich fühlt, wie man denkt, wie man sich verhält, wie erlebt man Gefühle und wie man sich bewegt. Diese Probleme können sehr lange bestehen und können nicht auf eine andere Krankheit oder auf den Gebrauch von Medikamenten zurückgeführt werden. Diese Krankheit kann erstmals auftreten oder kann schon mehrfach auftreten. Wenn sie erstmals auftreten, wird sie Schizophrenie, erste Episode genannt. Wenn sie mehrfach auftreten, wird sie Schizophrenie, mehrfache Episoden genannt. Die Symptome können während der Krankheit wieder zurückgehen oder verschwinden, aber sie werden nie</p>

Table 4: Examples of simplified texts in German, 6A20, run 9 (with increase in MTL D by 91.69 and decrease in FRE by 10.38) and run 7 (with decrease in MTL D by 37.55 and increase in FRE by 39.87). These are two contrastive examples: 1) complex words and no repetitiveness, dense enumeration, too complex; 2) simpler words but high repetitiveness, too simplified. The truncated sentences at the end are due to limit in tokens number; they were not manually edited.

Simplified text (fr, BioMistral, run 3)	Simplified text (fr, BioMistral, run 4)
<p>La schizophrénie est une maladie mentale où plusieurs types de pensées, perceptions, sentiments et comportements sont perturbés. Par exemple, quelqu'un peut avoir des idées délirantes (comme croire qu'une personne célèbre est en amour avec lui), des hallucinations (comme voir ou entendre quelqu'un qui n'est pas réellement là), une impression que ses propres sentiments, impulsions, pensées ou comportements sont sous le contrôle d'une force externe, un trouble de l'attention, de la mémoire verbale et de la cognition sociale, une perte de motivation, une expression émotionnelle éteinte, un comportement qui semble bizarre ou inutile, des réactions émotionnelles imprévisibles ou inappropriées qui</p>	<p>La schizophrénie est une maladie de l'esprit qui affecte plusieurs facettes de la pensée, de la perception, de l'expérience de soi, de la cognition, de la volonté, de l'affect et du comportement. Les symptômes fondamentaux sont des idées délirantes persistantes, des hallucinations persistantes, des troubles de la pensée et des expériences d'influence, de passivité ou de contrôle. Pour être diagnostiquée, les symptômes doivent persister pendant au moins un mois et ne doivent pas être dus à une autre maladie ou à l'effet d'une substance ou d'un médicament. La schizophrénie est une maladie qui peut avoir plusieurs épisodes, avec des périodes de rémission entre les é</p>

Table 5: Examples of simplified texts in French, 6A20, run 3 (with increase in MTL D by 43.51 and decrease in FRE by 23.74) and run 4 (with decrease in MTL D by 29.92 and increase in FRE by 5.68). These are two contrastive examples: 1) complex words and no repetitiveness, dense enumeration, too complex; 2) a bit simpler words (not much difference with original text) but high repetitiveness, too simplified. The truncated sentences at the end are due to limit in tokens number; they were not manually edited.