Patient-friendly Clinical Notes: Towards a new Text Simplification Dataset

Jan Trienes[†] **Jörg Schlötterer**[†] **Hans-Ulrich Schildhaus**[‡] **Christin Seifert**[†]
[†]University of Duisburg-Essen, Germany
[‡]University Hospital Essen, Germany
[‡]Discovery Life Sciences, Kassel, Germany
{jan.trienes, joerg.schloetterer, christin.seifert}@uni-due.de
hans-ulrich.schildhaus@dls.com

Abstract

Automatic text simplification can help patients to better understand their own clinical notes. A major hurdle for the development of clinical text simplification methods is the lack of high quality resources. We report ongoing efforts in creating a parallel dataset of professionally simplified clinical notes. Currently, this corpus consists of 851 document-level simplifications of German pathology reports. We highlight characteristics of this dataset and establish first baselines for paragraph-level simplification.

1 Introduction

Many hospitals worldwide give patients access to their own clinical notes with the goal to strengthen patient autonomy and increase transparency of the care process (Delbanco et al., 2012). Yet, clinical notes are seldomly written with the patient in mind. Being a communication tool for doctors, clinical notes must use a precise and unambiguous medical vocabulary. With limited health literacy, these notes are therefore practically inaccessible to most patients (Sørensen et al., 2015). The urgency of making clinical notes accessible to patients is underlined by initiatives like *"What's my diagnosis?"* where medical doctors and students volunteer to translate patient notes into a simple language (Bittner et al., 2015). Approaches to automatic text simplification (TS) have the potential to assist with this time consuming manual process (Shardlow, 2014; Alva-Manchego et al., 2020).

However, there is a lack of resources to develop TS methods for clinical notes. Most commonly used resources for TS include, on the one hand, professionally simplified news articles such as Newsela (Xu et al., 2015) and OneStopEnglish (Vajjala and Lučić, 2018), and on the other hand, large scale but potentially noisy alignments of Wikipedia (Zhu et al., 2010; Jiang et al., 2020). In the medical domain, datasets cover consumer health lexicons (Cao et al., 2020), laymen summaries of scientific articles (Devaraj et al., 2021) and medical subsets of Wikipedia (Grabar and Cardon, 2018; van den Bercken et al., 2019; Van et al., 2020). In addition, there is a lack of parallel document-level TS datasets (with the notable exception of Newsela and OneStopEnglish). This makes it difficult to study document-level phenomena of TS such as sentence reordering, deletion and explanation generation (Alva-Manchego et al., 2019b; Zhong et al., 2020; Srikanth and Li, 2021).

Contributions. As a response to the lack of resources, we report initial results towards a parallel dataset of patient-friendly clinical notes. Currently, the dataset covers 851 German surgical pathology reports with document-level simplifications by medical professionals. We describe a lightweight simplification protocol that encourages an intuitive simplification style and medical correctness, while not demanding linguistic training on the side of the editors. Furthermore, we conduct first analyses of the dataset and establish baselines for automatic simplification. Overall, we find that the source text is dense, with partly incomplete sentences and a highly specialized vocabulary, while manually produced simplifications are on average longer with a more constrained vocabulary. We observe a variety of simplification operations in the references including lexical substitution, syntactic simplification, explanations and content selection. We believe that this dataset can be an interesting testbed for research on document-level TS.

Our goal with this dataset is to increase the diversity of resources that TS research has available in terms of the text types, domains and languages. Sharing this dataset with the research community while respecting ethical, legal and privacy standards is currently underway.[1]

[1]We share the experiment code and provide updates on the dataset at github.com/jantrienes/simple-patho.

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 19 - 27

Original (German)	Translation (English)
1. Klinische Angaben	**1. Background**
Original: Unklare Raumforderung proximaler Oberschenkel links. V.a. Lymphom, DD Weichteilsarkom. Entität?	**Original:** Undefined mass proximal thigh left. Suspicion of lymphoma, DDx soft tissue sarcoma. Entity?
Human Reference: Bei der vorliegenden Gewebeentnahme handelt es sich um eine unklare Gewebeneubildung des oberen linken Oberschenkels. Es wird die Frage nach einer Diagnose gestellt.	**Human Reference:** The present tissue sample is an undefined tissue growth of the upper left thigh. The question of a diagnosis is raised.
System Output (mBART): Anhand der histologischen Untersuchungen wurde die Diagnose gesichert und dabei geht es um eine unklare Raumforderung am oberen Anteil des linken Oberschenkels. Es besteht der Verdacht auf ein Lymphom, ein Weichteilsarkom. Ein Weichteilsarkom ist ein bösartiger Tumor des Weichteilgewebes, welcher Merkmale von Fettzellen aufweist bzw. aus diesen hervorgeht.	**System Output (mBART):** By means of histological examinations the diagnosis was confirmed and it is about an undefined mass on the upper part of the left thigh. There is a suspicion of a lymphoma, a soft tissue sarcoma. A soft tissue sarcoma is a malignant tumor of the soft tissue, which has characteristics of fat cells or develops from them.
R-1/2/L: 33.33/11.43/30.56, **BLEU:** 7.25, **SARI:** 39.28	
2. Makroskopie	**2. Macroscopy**
Original: Proximaler Oberschenkel ventral links: mehrteilig beige-braunes, teils gelbliches Gewebe von zusammengelegt 2 x 2 bis 0,3 cm.	**Original:** Proximal thigh ventral left: several fragments of beige-brown, partly yellowish tissue of 2 x 2 to 0.3 cm when put together.
Human Reference: Eine makroskopische Untersuchung wird mit bloßem Auge durchgeführt und gibt eine erste Einschätzung über krankhafte Veränderungen. Außerdem soll diese Untersuchung eine Auswahl von Abschnitten für eine genauere Untersuchung mithilfe des Mikroskops liefern. Es handelt sich um eine Gewebeentnahme der Vorderseite des linken Oberschenkels. Das Präparat besteht aus mehreren Teilen, die beige-braun bis gelblich gefärbt sind und zusammengelegt 2 x 2 x 0,3 cm messen.	**Human Reference:** A macroscopic examination is done with the naked eye and allows an initial assessment of pathological changes. In addition, this examination shall provide a selection of slices suitable for a more detailed examination under the microscope. It is a tissue sample of the front of the left thigh. The specimen consists of several fragments that are beige-brown to yellowish in color and measure 2 x 2 x 0.3 cm when put together.
System Output (mBART): Die makroskopische Beurteilung beinhaltet eine Untersuchung und Beurteilung von eingesandten Proben mit dem bloßen Auge. Zur Untersuchung steht ein mehrereteilig beige-braunes, teils gelbliches Gewebe von zusammengelegt 2 x 2 bis 0,3 cm.	**System Output (mBART):** The macroscopic assessment includes an examination and assessment of submitted specimens with the naked eye. The specimen to be examined is a several parts beige-brown, partly yellowish tissue of 2 x 2 to 0.3 cm when put together.
R-1/2/L: 34.86/9.35/31.19, **BLEU:** 3.92, **SARI:** 36.56	

Table 1: Example pathology report from the validation set including human reference and system output (mBART). Reports typically consist of four sections (background, macroscopy, microscopy and conclusion) and each section is one input for the paragraph-level simplification model. We color-code summarization/deletion, explanation and lexical simplification/paraphrasing. For each section, we also give the ROUGE, BLEU and SARI scores. The example is continued in Appendix Table 5.

2 Dataset Creation and Analysis

We describe our design decisions for the creation of a parallel corpus of clinical notes. An example report is given in Table 1.

2.1 Data Selection

We decided to focus on pathology reports of sarcoma patients since clinicians noted particularly high amounts of questions concerning these reports. Sarcomas are a rare type of cancer with many subtypes which can affect people of all ages. The pathology report describes an analysis of tumor tissue and establishes the main diagnosis.

We sample reports from the electronic health records of the University Hospital Essen, a large research hospital in Germany. Each year, about 60,000 pathology reports are written by the pathology department. We identify suitable reports based on clinical codings (ICD-O-M). A query for the period of January 2019 until August 2021 yielded 1,644 reports on sarcoma patients. All reports were fully anonymized and we received ethics approval from our institutional review board.[2]

2.2 Simplification Protocol

To create a parallel corpus of original and simplified clinical notes, we ask medical experts how they would *intuitively explain* a given report to a patient. We take a decidedly inductive approach here: while guidelines for simplified language exist,[3] it is not clear to what extent these are suitable for clinical notes, and if annotators without formal linguistic training could operationalize them. In the terminology of Allen (2009), we use an intuitive rather than a structural simplification process.

It is commonly accepted that a good simplification depends on the target audience (Xu et al., 2015; Bingel et al., 2018; Gooding, 2022). To better define the audience and ensure a common simplification goal among editors, we developed

[2]University of Duisburg-Essen; Reference: 21-10198-BO

[3]For example Basic English (Ogden, 1930); we refer to Saggion (2017) and Štajner (2021) for more examples.

	Document-Level		Paragraph-Level	
Statistic	Original	Simplified	Original	Simplified
Documents	851	851	3,280	3,280
Sentences	23,554	28,155	22,191	26,551
Tokens	327,466	462,994	299,365	433,027
Types	10,292	11,229	9,843	10,798
Words/doc	385	544	91	132
Words/sent	14	16	13	16
Avg. TTR	0.47	0.42	0.69	0.63
Avg. FRE	32.90	40.30	27.65	40.05
Novelty	63/84/91%		70/87/92%	
CMP	1.55		2.75	

Table 2: Statistics for a document-level and paragraph-level alignment of our dataset. TTR = type-token ratio, FRE = Flesch Reading-Ease, CMP = average compression. Novelty is the average percentage of 1/2/3-grams that appear in the simplified text but not in the original.

a *patient persona*. A persona is a rich description of a prototypical user of a software system, a tool often used in human-computer interaction research (Cooper, 1999). With this persona at hand, editors were asked *"What questions would this patient have about the report?"* Additionally, we provided following simplification guidelines to further increase consistency across editors: (i) preserve the section structure of the reports, (ii) use the same tense as the original report, and (iii) do not add any interpretations that go beyond the stated facts.

We hired a team of 9 medical students in their fourth year of studies. A senior pathologist provided guidance on clinical questions during regular meetings and through email. All reports were simplified by one editor at the document-level using a plain text editor with grammar and spellchecking functionality. We implemented several quality gates for consistency and medical correctness of simplifications. First, we used an initial trial period of 10 reports to refine the guidelines and to allow editors to get familiar with the task. Second, we held monthly meetings to discuss simplification challenges and examples. A chat platform was setup to resolve urgent questions in a timely manner. Over the span of one year, we simplified 851 reports with a total effort of 812 hours (median 50 min./report). The students were compensated for their work with 10.5€ per hour corresponding to the usual rate for student assistants in Germany.

2.3 Preprocessing

For studying the characteristics of our TS corpus, we apply minimal pre-processing. We segment each document into sentences and tokens using NLTK. To establish a reliable vocabulary size, we lemmatize the text using spaCy and replace tokens that only consist of digits, punctuation or combinations thereof with a special token.[4]

We found that most reports consist of four core sections: background, microscopy, macroscopy, and conclusion. Therefore, we also compile a section-aligned version of the dataset where we keep reports that have a one-to-one alignment for all core sections (820 out of 851 reports). This makes our dataset also amenable for paragraph-level simplification (Devaraj et al., 2021) in addition to document-level simplification.

2.4 Dataset Characteristics

To better characterize the dataset, we analyze several surface-level properties (see Table 2 for an overview). We focus on measures that were commonly reported in prior work (Xu et al., 2015; Dmitrieva and Tiedemann, 2021) including the number of sentences and tokens, the vocabulary size (types), the length of documents and sentences, and the n-gram novelty. The type-token ratio (TTR) is used as a measure of lexical diversity and the Flesch Reading-Ease (FRE, Flesch, 1948) serves as a first indication of changes in readability.[5]

Simplifications are on average 41% longer than the original text (Table 2). Through manual inspection, we identified two potential reasons. First, the original reports tend to use a brief writing style with partly incomplete sentences. These were expanded to full sentences by the editors. Second, editors often added contextual information and explanations (e.g., why an examination was done, and what the result mean to a patient). The most striking difference in length can be observed for the background section (Figure 1). We assume that simplifications are "setting the scene" in this section by simplifying terminology and explaining concepts which do not have to be repeated again in the remainder of the report.

Simplifications select and summarize content. While simplifications are longer than their original counterpart, we also note a form of summarization (see example in Table 1). In some cases, particularly technical concepts were not included in the simplification, presumably because there is no simple explanation or because an explanation

[4] `nltk.org` and `spacy.io`

[5] We use constants adapted to German text (Amstad, 1978). Implementation in `github.com/textstat/textstat`.

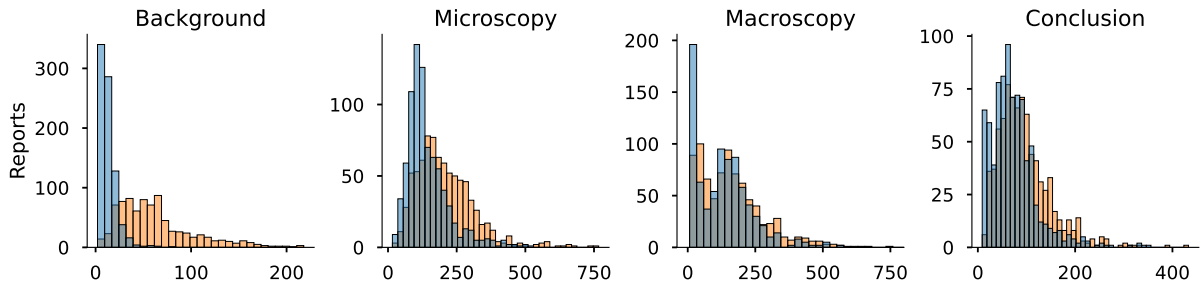

Figure 1: Comparing section length in the number of tokens for the Original and Simplified text. We observe largest expansion in the Background section. Simplifications for other sections follow the original length more closely.

would not help a user to better understand the report. This is in line with prior work which argues that document-level TS also requires summarization (Zhong et al., 2020; Aumiller and Gertz, 2022).

Simplifications have a different and more constrained vocabulary. While the simplified corpus is substantially larger in the number of tokens, the vocabulary size has only slightly increased. This is reflected in the lexical diversity measure (TTR: 0.47→0.42, Table 2). A decrease in TTR indicates that simplifications use a more constrained vocabulary which might translate to better readability. Furthermore, we observe a high average rate of unigram novelty (around 63%), which signals that large parts of the vocabulary are not shared.

Simplifications have a slightly higher readability. We observe a small increase in the readability measure (FRE: 32.9→40.3, Table 2). However, the overall readability is low according to this measure. By means of comparison, Aumiller and Gertz (2022) reported FRE values of 40 for the original and 67 for the simplified parts of a document-level TS dataset collected from German Wikipedia. There are inherent limitations with readability measures like FRE, so this finding has to be interpreted with care (Tanprasert and Kauchak, 2021).

3 Simplification Baselines

We next establish a first baseline for pathology report simplification using paragraph-level sequence-to-sequence methods (Devaraj et al., 2021).

3.1 Modeling Considerations

As discussed in Section 2.4, our dataset features multiple simplification operations including lexical simplification, paraphrasing, summarization and explanation generation. Therefore, we focus on monolingual neural machine translation models which can learn these operations simultaneously (Nisioi et al., 2017). Prior work on medical text investigated lexical simplification (Abrahamsson et al., 2014; Kloehn et al., 2018) or hybrid systems that combine pre-trained translation models with domain-specific phrase tables (Shardlow and Nawaz, 2019). With our parallel dataset, fine-tuning large general-purpose language models becomes a realistic option (Rothe et al., 2020).

Inspired by Devaraj et al. (2021), we train a paragraph-level simplification model. Compared with sentence-level methods, a paragraph-level model has the benefit that we do not need sentence alignments (Štajner et al., 2018) and that we can capture simplification phenomena like syntactic simplification and summarization (Alva-Manchego et al., 2019b). Our dataset has a natural paragraph-level alignment in the form of four core sections, so we consider this a suitable first baseline.

Methods. We experiment with four instantiations of paragraph-level methods. (1) **Identity**: A simple baseline which outputs the original text as simplification. (2) **Bert2Bert**: A transformer-based encoder-decoder where both parts are initialized with BERT (Devlin et al., 2019; Rothe et al., 2020). (3) **Bert2Share**: Same as Bert2Bert, but weights of the encoder and decoder are shared. (4) **mBART**: A sequence-to-sequence transformer, pre-trained on a sentence reconstruction objective (Liu et al., 2020). We include hyperparameters and replication details in Appendix A.

Evaluation. We report the standard TS metrics SARI (Xu et al., 2016), BLEU (Papineni et al., 2002) and ROUGE F_1 (Lin, 2004) for unigram (R-1) and bigram (R-2) matches, and the longest common subsequence between the reference and system output (R-L). To calculate SARI and BLEU, we use the implementation in EASSE (Alva-Manchego et al., 2019a) with default settings. For ROUGE,

Model	R-1	R-2	R-L	BLEU	SARI	Len.	Nov.
Identity	29.6	14.3	28.6	10.8	11.2	92	0%
Bert2Bert	26.5	8.3	25.0	7.3	41.4	103	79%
Bert2Share	28.3	9.5	26.6	8.2	42.7	102	78%
mBART	**35.2**	**15.3**	**33.4**	**14.2**	**46.2**	129	65%

Table 3: Automatic simplification results on paragraph-aligned data. The identity baseline simply returns the input as simplification. For the reference simplification, the average length (Len.) is 132 tokens and the average unigram novelty (Nov.) is 70% (cf. Table 2).

	Identity			mBART		
Section	BLEU	SARI	Len.	BLEU	SARI	Len.
Background	0.1	6.2	15	6.5	47.9	86
Macroscopy	17.2	12.5	136	18.0	48.2	131
Microscopy	8.7	10.7	146	13.0	44.3	213
Conclusion	13.9	10.5	72	13.6	43.6	88
Micro Avg.	10.8	11.2	92	14.2	46.2	129

Table 4: Evaluation by report section. Micro averaged metrics over all sections are reproduced from Table 3.

we use the rouge-score package with stemming disabled. We randomly split reports into training/validation/test sets with an 80/10/10 ratio.

3.2 Results and Discussion

Quantitative Results. According to automatic metrics, the generated simplifications have a substantially higher simplicity (SARI) but only slightly higher adequacy (ROUGE and BLEU) than an identity baseline (Table 3). mBART provides best results with an average simplification length and novelty close to the reference (129 vs. 132 tokens, and 65% vs. 70% novelty, Table 3). While not directly comparable, metrics are in a similar range as the paragraph-level simplification results on English medical abstracts by Devaraj et al. (2021).

For a better intuition of where the models can be improved, we report metrics by section type in Table 4. We see that the background section is most difficult to simplify. The low BLEU score of the identity baseline (0.1 in Table 4) indicates that there is little overlap between the original and simplified vocabulary. We hypothesize that simplifications for the background section include explanations and contextual domain knowledge which are difficult to generate with sequence-to-sequence methods (Srikanth and Li, 2021).

Qualitative Observations. By manual inspection, we found that system outputs are mostly fluent, grammatical and subjectively easier to read (Table 1). Furthermore, we observe that models generate elaborations and perform a certain degree of content selection. We also found factual errors in the automatically generated simplifications. In the example in Table 5, a clinical result was reported as positive in the original report but negative in the generated simplification (STAT6 positive vs. negative). The subsequently generated sentence (*"This combination of tumor markers is suggestive of GIST"*) is a clinically conceivable statement, but in the context of this report wrong. We anticipate that factual correctness will be of high importance for any practical deployment of a TS system for clinical notes and consider the evaluation of factual correctness as a significant avenue for future work on this dataset (Devaraj et al., 2022).

4 Conclusion and Future Work

We present ongoing work towards a dataset of professionally simplified clinical notes. Currently, the corpus consists of 851 parallel documents totaling close to 790k tokens. Quantitative and qualitative analyses show potential challenges for paragraph-level and document-level TS research. Despite a moderately sized training set, fine-tuning general language models led to promising results.

In future work, we will increase the size of the dataset and conduct a formal analysis of the simplification operations in the data to better understand the challenges for TS on clinical notes. Human evaluations with a focus on factual correctness, as well as user studies with end-users such as patients and patient advocacy groups are also envisioned.

Acknowledgements

We thank Celina Bandowski, Theresa Bernsmann, Monika Coers, Lisa Gödde, Hannah Göke, Lisa Meißner, Ral Merjanah, Justin Roschlak and Chiara Wedekind for writing the simplifications. We also thank Andrea Tonk for helping to translate the example report into English.

References

Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 57–65.

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019a. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 49–54.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019b. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A german dataset for joint summarization and simplification. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 2693–2701.

Joachim Bingel, Gustavo H. Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 245–258.

Anja Bittner, Johannes Bittner, and Ansgar Jonietz. 2015. *"Was hab' ich?" Makes Medical Specialist Language Understandable for Patients*, pages 331–338. Springer International Publishing.

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 (ACL)*, pages 1061–1071.

Alan Cooper. 1999. The inmates are running the asylum. In *Software-Ergonomie '99*, volume 53 of *Berichte des German Chapter of the ACM*.

Tom Delbanco, Jan Walker, Sigall K. Bell, Jonathan D. Darer, Joann G. Elmore, Nadine Farag, Henry J. Feldman, Roanne Mejilla, Long Ngo, James D. Ralston, Stephen E. Ross, Neha Trivedi, Elisabeth Vodicka, and Suzanne G. Leveille. 2012. Inviting patients to read their doctors' notes: A quasi-experimental study and a look ahead. *Annals of internal medicine*, 157(7):461–470.

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

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7331–7345.

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 (NAACL)*, pages 4171–4186.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 50–57.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9.

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 (ACL)*, pages 7943–7960.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Nicholas Kloehn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P. Yuan, and Debra Revere. 2018. Improving consumer understanding of medical text: Development and validation of a new SubSimplify algorithm to automatically generate term explanations in English and Spanish. *Journal of Medical Internet Research*, 20(8).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–91.

Charles Kay Ogden. 1930. Basic English: A general introduction with rules and grammar.

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 (ACL)*, pages 311–318.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(1).

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 380–389.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 5123–5137.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 2637–2652.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC)*.

Kristine Sørensen, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, Barbara Kondilis, Demosthenes Agrafiotis, Ellen Uiters, Maria Falcon, Monika Mensing, Kancho Tchamov, Stephan van den Broucke, and on behalf of the HLS-EU Consortium Brand, Helmut. 2015. Health literacy in europe: Comparative results of the european health literacy survey (HLS-EU). *European Journal of Public Health*, 25(6):1053–1058.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 1–14.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: The autocomplete for medical text simplification. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 1424–1434.

Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference (WWW)*, pages 3286–3292.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45.

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.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9709–9716.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1353–1361.

A Implementation Details

Hyperparameters. All simplification models were trained for 25 epochs using the AdamW optimizer with an initial learning rate of $3e{-}5$ (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). We use a learning rate schedule with an initial warmup period of 10% of the training steps and a linear decay afterwards. Checkpoints are taken every epoch and the checkpoint with lowest validation loss is kept. For Bert2Bert and Bert2Share, we set the batch size to 16 and for mBART to 4. During inference, we use beam search decoding with 5 beams. Generation ends when an end-of-sequence token is generated. We did not perform any manual or automatic hyperparameter tuning.

Implementation. To adapt mBART for the task of monolingual translation, we follow recommendations by Rios et al. (2021) and add a special language token for the original text and for the simplified text. We implement the models using the Transformers library (Wolf et al., 2020). Models are initialized with the bert-base-multilingual-cased and facebook/mbart-large-cc25 checkpoints.

Computation Cost. All models are trained on a single NVIDIA RTX A6000 GPU with 48GB of memory. Training duration is 2:45h for mBART, 1:30h for Bert2Bert and 1:16h for Bert2Share.

Original (German)	Translation (English)
3. Mikroskopie	**3. Microscopy**
Original: Mikroskopisch zeigt die Biopsie Anteile eines spindelzellig gestalteten Tumors. Dieser zeigt ein relativ monomorphes Bild mit einem hämangioperizytomartigem Gefäßmuster. Die Tumorzellen besitzen vergrößerte, leicht vesikuläre Zellkerne. Mitosefiguren sind kaum erkennbar (1/10 HPF). Das Stroma ist relativ fein und enthält einzelne Kollagenfasern. Nekrosen sieht man nicht. Ergänzend wurden immunhistochemische Untersuchungen durchgeführt. Der Tumor zeigt eine kräftige Positivität für CD34 und eine kräftige nukleäre Expression von STAT6. Die folgenden Antigene werden vom Tumor nicht exprimiert: Aktin, Caldesmon, Panzytokeratin (CKplus), Desmin, EMA, MUC4, S100, SOX10 und TLE1.	**Original:** Microscopically, the biopsy shows portions of a spindle-cell shaped tumor. The tumor has a relatively monomorphic appearance, with a hemangiopericytoma-like vascular pattern. The tumor cells have enlarged, slightly vesicular nuclei. Mitotic figures are barely visible (1/10 HPF). The stroma is relatively fine and contains single collagen fibers. Necroses are not detectable. Additionally, immunohistochemical examinations were conducted. The tumor shows strong positivity for CD34 and strong nuclear expression of STAT6. The following antigens are not expressed by the tumor: Actin, caldesmon, pancytokeratin (CKplus), desmin, EMA, MUC4, S100, SOX10, and TLE1.
Human Reference: Nachdem die Gewebeproben in schmale Schnitte weiterverarbeitet wurden, können sie nach weiterer Aufarbeitung (z.B Färbung) unter dem Mikroskop betrachtet werden. Unter dem Mikroskop erkennt Anteile eines Tumors aus spindelförmigen Zellen. Der Tumor zeigt in sich ein recht gleichartiges Bild. Die Blutgefäße wachsen in einem speziellen Muster. Man erkennt viele kleine, verzweigte Gefäße. Die Tumorzellen weisen vergrößerte, leicht blasenförmige Zellkerne auf. Zellkerne sind der Ort in einer Zelle, in der das Erbgut in Form von DNA gespeichert wird. Mitosefiguren stellen unter dem Mikroskop sichtbare Chromosomenstrukturen dar, die während der Zellteilung auftreten. Damit geben Sie Aufschluss über die Teilungsfähigkeit der Tumorzellen. Sie kommen nur vereinzelt vor. Das die Zellen umgebende Gewebe ist fein und enthält einzelne Kollagenfasern. Abgestorbene Gewebebereiche sind nicht sichtbar. Die Schnitte der Gewebeproben wurden außerdem immunhistochemisch angefärbt. Dies bedeutet, dass spezielle Stoffe genutzt wurden, welche eine Farbreaktion auslösen, sobald diese an bestimmte Strukturen an der Oberfläche und im Inneren der Tumorzellen binden. Durch diese Anfärbemethode kann die Art des Tumors weiter bestimmt werden. Es wurde CD34 und STAT6 nachgewiesen. Die Kombination aus diesen beiden Markern ist ein Kriterium für die Diagnose eines solitären fibrösen Tumors. Dies bezeichnet einen bösartigen Tumor mit Bindegewebseigenschaften. Bösartige Tumore unterscheiden sich von gutartigen vor allem in der Fähigkeit Tochtertumore ausbilden zu können und über Gewebegrenzen hinaus zu wachsen.	**Human Reference:** After being processed into narrow slices and further preprocessing (e.g. staining), the tissue samples can be viewed under the microscope. Under the microscope, parts of a tumor consisting of spindle-shaped cells can be seen. The tumor itself shows a rather homogeneous picture. The blood vessels grow in a particular pattern. Many small, branched vessels are visible. The tumor cells show enlarged, slightly vesicular nuclei. Nuclei are the place in a cell where genetic material is stored as DNA. Mitotic figures are chromosome structures that appear during cell division and are visible under the microscope. They thus provide information about the ability of the tumor cells to divide. There is only an isolated number of them. The tissue surrounding the cells is fine and contains single collagen fibers. Dead tissue areas are not visible. Moreover, the slice of the tissue samples were immunohistochemically stained. This means that special substances were used which trigger a color reaction as soon as they bind to certain structures on the surface and inner parts of tumor cells. This staining method can further determine the type of tumor. CD34 and STAT6 were detected. The combination of these two markers is a criterion for the diagnosis of a solitary fibrous tumor. This is a malignant tumor characterized by connective tissue. Malignant tumors differ from benign ones mainly in their ability to develop secondary tumors and to grow beyond tissue boundaries.
System Output (mBART): Die präparierten Gewebeproben wurden zur Aufklärung der Gewebeart und der Ausbreitung des Tumors mikroskopisch untersucht. Es wurden zusätzlich Antikörperuntersuchungen (Immunhistochemische Untersuchungen) für die weitere Bestimmung des Tumors durchgeführt. In der mikroskopischen Untersuchung zeigen sich Anteile eines Tumors aus spindelförmigen Zellen. Die Tumorzellen sehen im Vergleich zu gesunden Zellen verändert aus und weichen von der Norm ab. Es sind nur wenige Zellen im Prozess der Zellteilung zu erkennen (1 Mitosefigur/ 10HPF), was für ein langsames Wachstum des Tumors spricht. Ergänzend wurden immunhistochemische Untersuchungen durchgeführt. Hierbei werden mithilfe von Antikörpern Zellbestandteile nachgewiesen, welche in Tumoren typischerweise vorkommen und somit Informationen über die Art und Herkunft des Tumors liefern. Von den getesteten Tumormarkern (CD34, STAT6, Aktin, Caldesmon, Panzytokeratin (CKplus), Desmin, EMA, MUC4, S100, SOX10 und TLE1) sind CD34 positiv ausgefallen und STAT6 negativ ausgefallen. Diese Kombination aus Tumormarkern spricht für das Vorliegen eines Gastrointestinalen Stromatumors (GIST) oder einer Metastase eines GIST.	**System Output (mBART):** The prepared tissue samples were examined microscopically to elucidate the type of tissue and the spread of the tumor. Additionally, antibody examinations (immunohistochemical examinations) were performed to further determine the tumor. A microscopic examination revealed portions of a tumor consisting of spindle-shaped cells. The tumor cells look altered compared to healthy cells and deviate from the norm. Only a small number of cells are in the process of cell division (1 mitotic figure/ 10HPF), suggesting that the tumor grows slowly. In addition, immunohistochemical examinations were performed. In these examinations, antibodies are used to detect cellular components that are typically found in tumors and thus provide information about the type and origin of the tumor. Of the tumor markers tested (CD34, STAT6, actin, caldesmon, pancytokeratin (CKplus), desmin, EMA, MUC4, S100, SOX10 and TLE1), CD34 was positive and STAT6 was negative. This combination of tumor markers is suggestive of the presence of a gastrointestinal stromal tumor (GIST) or a metastasis of GIST.
R-1/2/L: 30.92/9.02/27.93, **BLEU:** 5.12, **SARI:** 35.94	
4. Epikritische Stellungnahme	**4. Conclusion**
Original: PE (proximaler Oberschenkel ventral links) mit einem solitären fibrösen Tumor (SFT), klassischer Typ. Zur Validierung des Befundes wurde ergänzend eine molekularpathologische Untersuchung (Fusionspanel) eingeleitet. Diesbezüglich folgt ein Nachbericht.	**Original:** BX (proximal thigh ventral left) with a solitary fibrous tumor (SFT), classic type. A supplementary molecular pathological examination (fusion panel) was initiated to validate the findings. There will be a follow-up report on this topic.
Human Reference: Es handelt sich um eine Probenentnahme der Oberschenkelvorderseite mit einem solitären fibrösen Tumor. Zum Nachweis dieser Diagnose wurden weitere Untersuchungen eingeleitet. Die Ergebnisse werden in einem Nachbericht geliefert.	**Human Reference:** This is a tissue sample of the thigh front with a solitary fibrous tumor. Further examinations were initiated to confirm this diagnosis. The results will be provided in a follow-up report.
System Output (mBART): Anhand der histologischen Untersuchungen wurde die Diagnose gesichert und dabei geht es um einen solitären fibrösen Tumor (SFT), klassischer Typ. Zur Validierung des Befundes wurde ergänzend eine molekularpathologische Untersuchung (Fusionspanel) eingeleitet. Diesbezüglich folgt ein Nachbericht.	**System Output (mBART):** By means of histological examinations the diagnosis was confirmed and it is about a solitary fibrous tumor (SFT), classic type. A supplementary molecular pathological examination (fusion panel) was initiated to validate the findings. There will be a follow-up report on this topic.
R-1/2/L: 40.58/11.94/34.78, **BLEU:** 5.07, **SARI:** 20.59	

Table 5: Example report (continued from Table 1).