A New Dataset and Efficient Baselines for Document-level Text Simplification in German

Anonymous EMNLP submission

Abstract

The task of document-level text simplification is very similar to summarization with the additional difficulty of reducing complexity. We introduce a newly collected data set of German texts, collected from the Swiss news magazine *20 Minuten* ('20 Minutes') that consists of full articles paired with simplified summaries. Furthermore, we present experiments on ATS with the pretrained multilingual mBART and a modified version thereof that is more memory-friendly, using both our new data set and existing simplification corpora. Our modifications of mBART let us train at a lower memory cost without much loss in performance, in fact, the smaller mBART even improves over the standard model in a setting with multiple simplification levels.

1 Introduction

Text simplification is the process of reducing the complexity of a text to make it more easily understandable by its readers and improve its accessibility for a wider audience. Depending on the use case, target groups of simplified texts may include low-proficiency readers such as persons with intellectual disabilities, prelingually deaf persons, or non-native readers. Automatic text simplification (ATS) employs natural language processing methods for generating a simplified version of a given text in standard language.

In general, simplification often results in a reduction of content similar to summarization, but with additional syntactic and lexical changes. Compression ratio in terms of sentence length or word count can be somewhat misleading since the simplified documents often elaborate on concepts and split complex sentences into smaller units.

Research on text simplification for German is still sparse but has gained momentum in recent years due to a number of legal and political developments in German-speaking countries, such as the introduction of a set of regulations for accessible information technology (*Barrierefreie-Informationstechnik-Verordnung*, BITV 2.0) in Germany, the approval of rules for accessible information and communication (*Barrierefreie Information und Kommunikation*, BIK) in Austria, and the ratification of the United Nations Convention on the Rights of Persons with Disabilities (UN CRPD) in Germany, Austria, and Switzerland.

In this work, we report on two contributions regarding ATS for German:

1. We introduce a new data set of simplified news articles from the Swiss daily magazine *20 Minuten* ('20 Minutes'). The source side of the corpus contains the full, standard German news, whereas the target side consists of a shortened and simplified version that is meant to give readers an easy and fast-to-read overview. This data set will be released upon publication.

2. We apply an adapted version of the mBART model (Liu et al., 2020) to the task of document-level ATS. The model needs to learn to reduce the content of the original document to the most salient parts, just as in summarization tasks. However, on top of that, the model also needs to account for linguistic changes that correspond to the targeted simplification level. All code will be made available upon publication.

In addition to the new *20 Minuten* data set, we evaluate our adapted mBART model with pre-existing corpora for German ATS (see Section 3).

2 Related Work

Traditionally, ATS has relied on rule-based approaches in separate steps, e.g. lexical substitutions followed by syntactic modifications. In the case of lexical simplification (i.e. the identification of

difficult words and the substitution with simpler synonyms), most modern approaches include features based on semantics, context, and language models (Glavaš and Štajner, 2015; Qiang et al., 2020). Syntactic simplification (the identification and simplification of difficult syntactic structures) is mostly done using manually written rules applied to a syntax tree (Siddharthan, 2006; Scarton et al., 2017). Such systems are still among the most successful for languages with little simplification data such as Basque (Aranzabe et al., 2012), Bulgarian (Lozanova et al., 2013), or French (Brouwers et al., 2014).

For languages with enough parallel data (i.e. mainly English), data-driven approaches that rely on machine learning have emerged, where ATS is most often framed as a monolingual machine translation task. Statistical machine translation has been applied to learn complex-simple phrase correspondences from parallel sentence-aligned corpora (Wubben et al., 2012), sometimes in conjunction with rule-based simplification (Narayan and Gardent, 2014) or via integration of syntactic information through syntax-based SMT (Xu et al., 2016a).

More recently, neural machine translation (NMT) has been used to train models to directly map complex to simple sentences. Supervised learning with recurrent or transformer architectures dominate current state-of-the-art research, some with additional simplification-specific adaptations such as lexical constraints, rule-based preprocessing, or parametrization mechanisms (Nisioi et al., 2017; Zhang and Lapata, 2017; Sulem et al., 2018; Mallinson and Lapata, 2019; Kriz et al., 2019; Martin et al., 2020a). Some unsupervised or semi-supervised neural models, which reduce the need for parallel data, have reached similar performances (Surya et al., 2019; Kumar et al., 2020; Zhao et al., 2020; Martin et al., 2020b). Finally, experiments with multi-task learning have shown promising results (Guo et al., 2018; Dmitrieva and Tiedemann, 2021), with the possibility of zero-shot translations for languages without any parallel data (Mallinson et al., 2020). These approaches represent the current state of the art, but are largely limited to English (Al-Thanyyan and Azmi, 2021) due to a lack of training data in other languages. Initial experiments with German are ongoing (Battisti et al., 2020).

Simplification operations often occur across sentence borders and can affect the structure of a text as a whole, which complicates sentence alignment and limits the effectiveness of sentence simplification models. Initial experiments exist that use document-level data to avoid these problems (Zhong et al., 2020; Dmitrieva and Tiedemann, 2021).

In this paper, we treat text simplification as a document-level task similar to summarization: the model needs to identify the most relevant information from the original text and generate a condensed version thereof. On top of that, the model should ideally learn to modify syntactic structures (e.g. split long sentences) and replace complex words (e.g. compound nouns) with simpler alternatives.

3 Data

We introduce a new data set collected from the Swiss news magazine *20 Minuten* that consists of full articles paired with shortened, simplified summaries that serve as a quick "tl;dr" for the reader. In contrast to other data used in our work, this data set does not distinguish different simplification levels. The corpus contains a total of 18,305 articles published since 2020. For each article we collect the title, the lead, the full news text, and the summary. We also keep track of paragraph formatting, even though this information is not used in the models presented in this paper.

Additionally, we use a combination of two existing corpora for German ATS that explicitly label the difficulty level of the target documents according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2009). For some documents, we have multiple levels of simplification available.[1] The levels available to us are A1, A2 and B1 (from most simplified to close to standard German). The three corpora we use for our experiments have the following characteristics:

APA is an extended version of the Austrian Press Agency corpus described in Säuberli et al. (2020). This data set contains news articles professionally simplified to levels A2 and B1.

20m is a newly collected corpus from the Swiss news portal *20 Minuten*. Similar to the APA data,

[1]Note that our train/dev/test split is based on document IDs: if a document has multiple versions in different levels, we assign all of those to the same split, in order to avoid a scenario where we would train on a document de→A2 and then test on the same document with de→B1, as this would give the model an unfair advantage.

these are news articles paired with condensed, simplified summaries. The target side in this corpus does not distinguish between simplification levels.

capito is a corpus of documents from *capito*, the largest provider of human simplification services for German. This data set covers a wide range of topics and domains, from official information (e.g. what to do in case of a suspected covid infection) to local news, technical guidelines and instruction manuals. The *capito* documents are much more varied than the other data sets, both in content and length. The translations in this corpus include levels A1, A2 and B1.

For both the APA and the 20m data set, the compression ratio is comparable to summaries, as the simplified documents are generally much shorter than the original text. For the *capito* data, this is not always the case, at least in terms of word count; the simplified texts often elaborate on concepts or processes, which leads to a similar word count between the standard and the simplified documents. However, regarding content, the simplified texts usually do condense the original information to the most salient facts. For this reason, we argue that even on this data set, the task is very similar to summarization.

Table 1 illustrates the size of the different data sets and compression ratios according to simplification levels. See Appendix A.2 for samples from all three data sets.

4 Model and Training

Initial experiments showed that fine-tuning the standard pretrained mBART model (Liu et al., 2020) from Huggingface (Wolf et al., 2020), performs relatively well with our data, however, training is very memory-intensive, requiring a 32GB GPU even with a small batch size. For this reason, we modify the original model to allow us to train on devices with less memory.[2] Our modifications are based on the code for BART with Longformer attention by the Allen Institute for AI (Beltagy et al., 2020).[3]

As in the BART model with Longformer attention, we swap the standard attention in the mBART encoder for Longformer's windowed attention.[4] This allows us to extend the maximum input positions to 4096 to cover most of the documents in our data. The new positional embeddings are initialized with a copy of the pretrained embeddings of size 1024, as described in Beltagy et al. (2020). The decoder remains unchanged with a maximum sequence length of 1024.

Furthermore, we reduce the original mBART vocabulary from 250k to 20k, keeping only those subwords and their embeddings that are most relevant for German. [5] We apply the pretrained multilingual sentencepiece model to ∼4.5 million German sentences[6] and use the most frequent 20k subwords to filter the original mBART vocabulary.

We then extend the special language tokens with tags for the different simplification levels (e.g. "de_A1"). These are initialized with the pretrained embedding for the German language tag ("de_DE") and updated during fine-tuning. And lastly, we add the option to train and translate mixed batches with multiple target language labels.

We train our models with early stopping according to rougeL on a held-out validation set. The models converge after training for 2 to 5 days, the exact configuration and hyperparameters can be found in Appendix A.1. All models are trained on a single V100 GPU with the same accumulated batch size (60), but note that the standard mBART can only fit a batch size of 1 on the GPU, whereas our modified version can fit 4 samples in a batch and thus needs fewer accumulation steps.

5 Results

The results in Table 2 for the CEFR-labeled APA+capito data clearly show that with higher simplification levels, the task becomes harder: scores for both the standard mBART and our modified version ('small mBART') generally decrease with increasing distance to standard German.[7] The

[2] All models in this paper are trained on 32GB V100 GPUs for comparability to the baseline standard mBART, but with our modifications, we can load and fine-tune mBART on smaller GPUs (tested on a single 12GB Titan X).

[3] https://github.com/allenai/longformer

[4] We use a more recent version of both pytorch lightning and huggingface libraries and therefore have to make some changes, not only to the Longformer code, but also to the mBART model in huggingface itself. All code will be released upon publication.

[5] The step of trimming the embedding matrix is the most effective in reducing the size of the model and allowing it to be fine-tuned on smaller devices.

[6] Parts of the Common Crawl corpus 2019, News Commentary v15, Europarl v10 (all available from http://www.statmt.org/wmt20/translation-task.html) and our own data, see Section 3.

[7] Apart from BLEU and rougeL, we evaluate with SARI (Xu et al., 2016b), a metric introduced specifically for ATS.

	train			dev			test			compression ratio		
	capito	APA	20m	capito	APA	20m	capito	APA	20m	capito	APA	20m
A1	652	–	–	50	–	–	50	–	–	54%	–	–
A2	1708	2250	–	87	113	–	91	109	–	97%	23%	–
B1	1074	2302	–	56	144	–	65	135	–	98%	25%	–
simple	–	–	17905	–	–	200	–	–	200	–	–	11%

Table 1: Number of documents with compression ratio. APA and *capito* use simplification levels A2/B1 and A1/A2/B1, respectively. 20m does not distinguish between simplification levels (labeled as 'simple'). See Appendix A.2 for examples.

		APA+capito			20m
		A1	A2	B1	simple
rougeL	mBART	21.68	24.27	28.46	21.62
	small mBART	26.05	26.22	29.40	19.96
SARI	mBART	30.85	32.42	32.88	33.29
	small mBART	32.35	32.90	32.87	33.29
BLEU	mBART	6.31	8.91	13.15	7.47
	small mBART	8.25	10.02	14.40	6.29

Table 2: Results of automatic simplification with fine-tuned standard mBART and our modified, smaller version with longformer attention (small mBART). Since standard mBART does not have labels for simplification levels, target language is set to 'de_DE' for fine-tuning and evaluation. Decoding for all models is done with beam size=6 and length penalty=1.

mBART modifications to reduce memory-usage come at a small loss in performance according to rougeL and BLEU on the 20m data set. However, this smaller model with the additional language level tags outperfoms standard mBART on the APA+capito data set. Overall, the 20m articles are harder to simplify, since the compression ratio is relatively high (11%, see Table 1).

6 Conclusions

In this paper, we have introduced a data set of simplified news articles from the Swiss magazine *20 Minuten*, aligned on document level. The task of document-level simplification resembles that of summarization, as models need to identify the salient parts and produce a condensed version of the original text. For simplification, models should also learn to simplify syntactic structures and lexical items.

Experiments based on fine-tuning the pretrained mBART model from huggingface show that the model can learn to produce not just condensed, but also simpler output. Our added modifications make mBART fine-tuning significantly more memory-friendly. Since the new 20m data set does not distinguish between simplification levels, we use an existing data set annotated with CEFR levels (Säuberli et al., 2020) to evaluate our models according to specific simplification levels. Results show that our modified mBART, while using considerably less memory, can simplify documents without much loss in performance on the 20m data and even improves over standard mBART on documents labeled with CEFR tags.

In future work, we will conduct ablation studies to measure the effect of our modifications individually. We are interested in seeing whether e.g. using windowed attention to give the model access to the full source document instead of a clipped version is beneficial. Lastly, automatic evaluation with metrics such as rougeL, BLEU, and SARI do not provide sufficient insights. To get more accurate feedback and better understand issues specific to simplification, we plan to conduct an evaluation with professional human translators.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Marıa Jesús Aranzabe, Arantza Dıaz De Ilarraza, and Itziar Gonzalez-Dios. 2012. First approach to automatic text simplification in basque. In *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8, Istanbul, Turkey.

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 12th Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56, Gothenburg, Sweden. Association for Computational Linguistics.

Council of Europe. 2009. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Anna Dmitrieva and Jörg Tiedemann. 2021. A multitask learning approach to text simplification. *Recent Trends in Analysis of Images, Social Networks and Texts*, 1357:78.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. Association for Computational Linguistics.

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(0):726–742.

Slavina Lozanova, Ivelina Stoyanova, Svetlozara Leseva, Svetla Koeva, and Boian Savtchev. 2013. Text modification for Bulgarian Sign Language users. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 39–48, Sofia, Bulgaria. Association for Computational Linguistics.

Jonathan Mallinson and Mirella Lapata. 2019. Controllable sentence simplification: Employing syntactic and lexical constraints. *arXiv preprint arXiv:1910.04387*.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020a. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020b. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 85–91,

Vancouver, Canada. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 8649–8656. Association for the Advancement of Artificial Intelligence.

Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking data-driven automatic text simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 41–48, Marseille, France. European Language Resources Association.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan. Association for Computational Linguistics.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

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

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016b. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4(401–415).

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Xu Zhao, Zihao Wang, Hao Wu, and Yong Zhang. 2020. Semi-supervised bilingual lexicon induction with two-way interaction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2973–2984, Online. Association for Computational Linguistics.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 9709–9716. Association for the Advancement of Artificial Intelligence.

A Appendix

A.1 Model Configurations

	standard mBART	small mBART
max output length	1024	1024
max input length	**1024**	**4096**
batch size	**1**	**4**
gradient accumulation	**60**	**15**
gpus	1	1
seed	222	222
attention dropout	0.1	0.1
dropout	0.3	0.3
attention mode	–	**sliding chunks**
attention window size	–	**512**
label smoothing	0.2	0.2
learning rate	0.00003	0.00003
early stopping metric	rougeL	rougeL
patience	10	10
min delta	0.0005	0.0005
learning rate scheduler	ReduceOnPlateau	ReduceOnPlateau
lr reduce patience	8	8
lr reduce factor	0.5	0.5
vocabulary size	**250k**	**20k**

Table 3: Training configurations for standard mBART fine-tuning and modified version. Differences highlighted in bold.

A.2 Examples

Original (standard German)	English
Bestätigung über die Durchführung der Erstunterweisung	Confirmation that the initial instruction has been carried out
Ich habe die Erstunterweisung gemäß § 14 ASchG für neue MitarbeiterInnen erhalten.	I have received the initial instruction in accordance with § 14 ASchG for new employees.
Mit meiner Unterschrift auf diesem Blatt bestätige ich, dass ich die Information über:	With my signature on this sheet, I confirm that I have received and read information about:
atempo und Arbeitssicherheit	atempo and occupational safety
Gefahren durch Elektrizität	dangers due to electricity
Benützung von Leitern	use of ladders
Unfälle und Erste Hilfe	accidents and first aid
Wichtige Personen und Telefonnummern	important people and telephone numbers
gelesen und verstanden habe.	
A2 German	**English**
Bestätigung: Ich habe die wichtigen Informationen bekommen	Confirmation: I have received the important information
Sie müssen wichtige Informationen bekommen, wenn Sie in einer Firma anfangen.	You should receive important information when you start in a company.
Wenn Sie unterschreiben, heißt das: Sie haben die wichtigen Informationen gelesen und verstanden.	Your signature means: you have read and understood the important information.
Sie haben wichtige Informationen zu diesen Themen bekommen:	You have received important information on these topics:
Wie arbeite ich sicher?	How do I work safely?
Wie kann ich ein Feuer vermeiden?	How can I avoid a fire?
Was muss ich tun, wenn es brennt?	What should I do if there is a fire?
Warum kann Strom gefährlich sein?	Why can electricity be dangerous?
Wie muss ich eine Leiter benutzen?	How do I use a ladder?
Was muss ich tun, wenn sich jemand verletzt?	What should I do if someone is injured?
Wer sind wichtige Personen bei atempo?	Who are the important people at atempo?
Wo stehen wichtige Telefon-Nummern?	Where can I find the important phone numbers?
A1 German	**English**
Wichtige Informationen	Important Information
Sie sind neu bei atempo.	You are new at atempo.
Wir müssen Ihnen wichtige Informationen geben.	We have important information to give you.
Wie arbeite ich sicher?	How do I work safely?
Es brennt.	There is a fire.
Warum ist Strom gefährlich?	Why is electricity dangerous?
Wie muss ich eine Leiter verwenden?	How do I use a ladder?
Eine Person hat sich weh getan.	Someone is injured.
Wer sind wichtige Personen bei atempo?	Who are the important people at atempo?
Wo stehen wichtige Telefon-Nummern?	Where can I find the important phone numbers?
Haben Sie diese Informationen bekommen?	Did you receive this information?
Haben Sie diese Informationen verstanden?	Did you understand this information?
Dann unterschreiben Sie bitte diesen Zettel.	Then please sign this piece of paper.

Table 4: *capito* simplification example for levels A2 and A1 with elaborations. Document length for *capito* varies considerably, from documents with one sentence to documents with several thousand sentences.

Original (standard German)	English
"Lonely Planet" kürt Salzburg für 2020 zur besten Stadt Salzburg ist im kommenden Jahr für den Reisebuchverlag "Lonely Planet" die beste Stadt zum Bereisen. Im neuen "Lonely Planets Best in Travel 2020" führt die Mozartstadt das Ranking in der Kategorie der Städte nicht zuletzt wegen des 100-Jahr-Jubiläums der Festspiele an. Der Reiseführer "Best in Travel" kürt jedes Jahr zehn Top-Städte, -Länder und -Regionen. "Trommelwirbel, bitte", heißt es auf der Homepage des Verlages. "Der Herzensbrecher einer Alpenstadt besingt das Jubiläum in vollen Tönen." Salzburg führt das Ranking 2020 vor den Städten Washington DC, Kairo, dem irischen Galway und der Beethoven-Stadt Bonn an. In der Länderkategorie liegt Buthan voran, als Top-Region wurde die Seidenstraße in Zentralasien angegeben. Österreich kommt im Ranking 2020 kein zweites Mal vor.	"Lonely Planet" selects Salzburg as the best city to visit in 2020 Salzburg is the best city to travel to next year according to the travel book publisher "Lonely Planet". In the new "Lonely Planet's Best in Travel 2020", the city of Mozart leads the ranking in the category of cities, not least because of the 100th anniversary of the festival. The "Best in Travel" publication selects the top ten cities, countries and regions each year. "Drum roll, please," reads the publisher's homepage. "The heartbreaker of an Alpine city celebrates the anniversary in full tones." Salzburg leads the 2020 ranking ahead of Washington DC, Cairo, Galway, Ireland, and Bonn, the city of Beethoven. Buthan leads in the country category, with the Silk Road in Central Asia given as the top region. Austria does not appear a second time in the 2020 ranking.
B1 German	English
Reiseführer erklärt Salzburg zur besten Stadt der Welt Die österreichische Stadt Salzburg ist weltweit die beste Stadt zum Bereisen im kommenden Jahr. Das sagt die Rangliste des britischen Reiseführers "Lonely Planet". "Lonely Planet" erstellt jedes Jahr eine Rangliste der besten 10 Städte, Länder und Regionen auf der ganzen Welt. Für das Jahr 2020 liegt Salzburg auf Platz 1. Salzburg führt vor den Städten Washington in den USA, Kairo in Ägypten, Galway in Irland und Bonn in Deutschland. In der Rangliste der besten Länder zum Bereisen 2020 gewann das Land Buthan in Süd-Asien.	Travel guide declares Salzburg the best city in the world The Austrian city of Salzburg is the best city in the world to travel in the coming year. The ranking of the British travel guide "Lonely Planet" says so. "Lonely Planet" ranks the best 10 cities, countries and regions around the world each year. For the year 2020, Salzburg made it to first place. Salzburg leads, ahead of the following cities: Washington in the USA, Cairo in Egypt, Galway in Ireland and Bonn in Germany. In the ranking of the best countries to travel in 2020, the country of Buthan in South Asia won.
A2 German	English
Salzburg ist 2020 die beste Stadt zum Bereisen Die Stadt Salzburg ist im Jahr 2020 die beste Stadt zum Bereisen. Das sagt der Verlag von den Reise-Büchern namens Lonely Planet. Salzburg gewann vor den Städten Washington in den USA, Kairo in Ägypten und Galway in Irland. Der Verlag sucht jedes Jahr die besten 10 Städte, Länder und Regionen zum Bereisen.	Salzburg is the best city to travel in 2020 The city of Salzburg is the best city to travel in 2020. The publisher of travel books called Lonely Planet says so. Salzburg won ahead of the following cities: Washington in the USA, Cairo in Egypt and Galway in Ireland. Every year, the publisher looks for the best 10 cities, countries and regions to travel.

Table 5: APA example for levels A2 and B1. APA news articles are generally relatively short with up to ∼100 sentences.

Original (standard German)	English
Eine 58-jährige Frau war am Samstag, um 15.15 Uhr mit dem Auto auf der St. Gallerstrasse in Gossau unterwegs.	A 58-year-old woman was driving her car on St. Gallerstrasse in Gossau at 3:15 p.m. on Saturday.
Während der Fahrt bemerkte die Frau, dass sie ihr Handy auf dem Autodach vergessen hatte.	While driving, the woman noticed that she had forgotten her cell phone on the roof of the car.
Sie bremste ab.	She braked.
In diesem Moment fuhr ein 22-jähriger Töfffahrer hinter ihr.	At that moment, a 22-year-old motorcyclist was driving behind her.
Wie die Kantonspolizei St. Gallen mitteilt, war das Handy mittlerweile zu Boden gefallen und der Töfffahrer richtete seinen Blick auf den Gegenstand am Boden.	According to the cantonal police of St. Gallen, the cell phone had fallen to the ground and the motorcyclist turned his gaze to the object on the ground.
Dabei bemerkte der Mann nicht, dass das Auto vor ihm abbremst.	In doing so, the man did not notice that the car in front of him was slowing down.
Er prallte mit dem Töff in das Auto der 58-Jährigen.	He crashed his motorcycle into the car of the 58-year-old.
Dabei erlitt der Töfffahrer unbestimmte Verletzungen.	The driver of the motorcycle suffered unspecified injuries.
Mit einem Rettungswagen wurde er ins Spital gebracht.	He was taken to hospital in an ambulance.
Laut der Polizei entstand ein Sachschaden von mehr als 20"000 Franken.	According to the police, the damage to property amounted to more than 20,000 Swiss francs.
Simplified German	**English**
Eine Autofahrerin hat ihr Handy auf dem Dach vergessen.	A car driver forgot her cell phone on the roof.
Als sie das bemerkte, bremste sie während der Fahrt ab.	When she realized, she braked abruptly while driving.
Ein Töfffahrer hinter ihr war durch das heruntergefallene Handy abgelenkt und prallte darauf in das Auto.	A motorcyclist behind her got distracted by the dropped cell phone and crashed into the car.
Der 22-jährige Töfffahrer erlitt unbestimmte Verletzungen.	The 22-year-old motorcyclist suffered unspecified injuries.

Table 6: 20m example, language levels are not distinguished in this corpus.