Towards Automatic Finnish Text Simplification

Anna Dmitrieva, Jörg Tiedemann
University of Helsinki
{name.surname}@helsinki.fi

Abstract

Automatic text simplification (ATS/TS) models typically require substantial parallel training data. This paper describes our work on expanding the Finnish-Easy Finnish parallel corpus and making baseline simplification models. We discuss different approaches to document and sentence alignment. After finding the optimal alignment methodologies, we increase the amount of document-aligned data 6.5 times and add a sentence-aligned version of the dataset consisting of more than twelve thousand sentence pairs. Using sentence-aligned data, we fine-tune two models for text simplification. The first is mBART, a sequence-to-sequence denoising auto-encoder proven to show good results for monolingual translation tasks. The second is the Finnish GPT model, for which we utilize instruction fine-tuning. This work is the first attempt to create simplification models for Finnish using monolingual parallel data in this language. The data has been deposited in the Finnish Language Bank (Kielipankki) and is available for non-commercial use, and the models are accessible through Huggingface.

Keywords: automatic text simplification, parallel dataset, Finnish

1. Introduction

In recent years, the number of non-English text simplification corpora has grown significantly. For example, there exist a number of simplification datasets for other European languages such as French (see, for example, Alector (Gala et al., 2020), CLEAR (Grabar and Cardon, 2018)), German (see: Klexicon (Aumiller and Gertz, 2022), Patient-friendly Clinical Notes (Trienes et al., 2022)), Italian (see: AdminIT (Miliani et al., 2022)), and others (more examples can be found in Ryan et al., 2023). It is worth noting that the past decade saw a growing movement toward media accessibility in European countries, including legal action such as implementing the Directive EU 2016/2102[1] on the accessibility of the websites and mobile applications of public sector bodies. This is one of the reasons why the interest in accessible communication studies for European languages other than English has increased.

In Finland, Easy Language is well-established in practice (Leskelä, 2021), and Easy Language content such as news, books, and websites is produced regularly. Nevertheless, the first parallel Finnish-Easy Finnish dataset (Dmitrieva et al., 2022) has been introduced only very recently (Dmitrieva and Konovalova, 2023). This dataset, however, is rather small, with only 1919 entries, and aligned only on the document level. In this work, we increase the size of this dataset by adding more aligned document pairs and producing a sentence-aligned version. Information on the base dataset can be found in Section 3, and our work on document and sentence alignment is described in Section 4. We then train different sentence simplification models to provide a baseline for automatic Finnish text simplification. Modeling is described in Section 5.

2. Related work

Using news as a data source is a popular approach to building simplification corpora for languages that have simplified news sources. We will name just a few examples. For instance, Ebling et al. (2022) describe a dataset consisting of articles from the Austria Press Agency (Austria Presse Agentur, APA). At this press agency, four to six news items covering the topics of politics, economy, culture, and sports are manually simplified into two language levels, B1 and A2, each day (Ebling et al., 2022). Rios et al. (2021) describe another parallel German simplification dataset based on news articles 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. Goto et al. (2015) describe a data set consisting of Japanese news sentences and their corresponding simplified Japanese news sentences sourced from a web resource called NEWS WEB EASY (Tanaka et al., 2013) offered by the NHK [Japan Broadcasting Corporation]. Finally, Newsela (Xu et al., 2015), a well-known simplification dataset with simplifications for four different grade levels, available in English and Spanish, is also news-based.

Since simplification can be viewed as a monolingual translation problem, researchers sometimes use tools intended for multilingual alignment of machine translation corpora to align monolingual simplification data. For example, Spring et al. (2023) use Vecalign (Thompson and Koehn, 2019) among other sentence aligners to analyze alignment qual-

[1] http://data.europa.eu/eli/dir/2016/2102/oj

21 May, 2024.

ity for automatic simplification of German texts, and Stodden et al. (2023) experiment with Vecalign and Bertalign (Liu and Zhu, 2022) to develop a new parallel dataset for German simplification. Vecalign also includes a tool that can be used for document alignment (Thompson and Koehn, 2020). Most of the alignment strategies require pre-trained embeddings, which can also be utilized on their own for parallel text detection (Spring et al., 2023; Stodden et al., 2023; Aumiller and Gertz, 2022). Specialized tools for monolingual alignment, such as MASSAlign (Paetzold et al., 2017) and CATS (Customized Alignment for Text Simplification) (Štajner et al., 2018), are also used for alignment, often in conjunction with other methods (see, for instance, Ebling et al., 2022).

In this work, we use two different architectures to create simplification models. BART (Lewis et al., 2020) models, including multilingual BART/mBART (Liu et al., 2020), are widely used for automatic text simplification and have shown good results for English (Martin et al., 2022), German (Trienes et al., 2022; Stodden et al., 2023), Spanish (Alarcón et al., 2023), and other languages. GPT models are used for simplification less often but still have shown good results, for example, for English (Maddela et al., 2023) and Russian (Shatilov and Rey, 2021). We use a GPT model trained on multiple Finnish resources (Luukkonen et al., 2023).

3. Data

We use three datasets as sources for our research: the Parallel Corpus of Finnish and Easy-to-read Finnish (Dmitrieva et al., 2022), the Yle Finnish News Archive 2011-2018 (Yleisradio, 2017) which we call the "general" archive because it consists of all news that appeared on yle.fi during these years, and Yle News Archive Easy-to-read Finnish 2011-2018 (Yleisradio, 2019). All of these datasets are available in the Language Bank of Finland [Kielipankki] under the CLARIN ACA-NC license (Academic - Non-Commercial Use, Attribution, No Redistribution, Other). The first parallel dataset is based on Yle articles from 2019 to 2020, so we are using articles from earlier times to increase the amount of parallel data.

YLE news in Easy Finnish comes on air every day in the form of short (around 5 minutes) radio and TV broadcasts relaying the most important recent events. The radio broadcast then appears on YLE's website in the form of an article, where each paragraph details its own piece of news. The target audience of Easy Finnish news is very broad, with the main target groups being immigrants, older adults, and people with intellectual disabilities (Kulkki-Nieminen, 2010).

The editors at YLE choose the material to simplify for Easy Finnish news themselves. There is no time frame for how recent the "regular" Finnish article should be, but the editors mostly select articles that came out in the 24 hours before the Easy Finnish broadcast airs (Dmitrieva and Konovalova, 2023). Therefore, for document alignment, we enforce the same limitation as Dmitrieva and Konovalova (2023) did in the original dataset and only align Standard Finnish and Easy Finnish documents from the same date. Unfortunately, we could not match articles prior to September 2014. Easy Finnish articles from before this date are mixed into the general news archive without any clear identifiers. Therefore, in this paper, we are working with articles from September 2014 to December 2020. We leave the identification of earlier Easy Finnish news in the general archive for future work.

4. Dataset augmentation

In this work, we first align more Standard and Easy Finnish articles and then produce a sentence-aligned version of the entire dataset. For both tasks, we use embedding models to produce document and sentence vectors. Here is the complete list of the embeddings that we use:

1. LASER[2] (we used the laserembeddings library[3]),
2. LaBSE (Feng et al., 2022; we used the version from sentence-transformers[4]),
3. MPNet (Song et al., 2020; we used the version from sentence-transformers[5]),
4. DistilUSE (multilingual knowledge distilled version of multilingual Universal Sentence Encoder (Yang et al., 2020)), also from sentence-transformers[6].

Three of these four models are multilingual sentence-BERT networks (Reimers and Gurevych, 2019). We have selected DistilUSE as a benchmark since it has been used in the making of the original dataset (Dmitrieva and Konovalova, 2023). We also chose MPNet because it has shown

[2] https://github.com/facebookresearch/LASER

[3] https://github.com/yannvgn/laserembeddings

[4] https://huggingface.co/sentence-transformers/LaBSE

[5] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[6] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

the best average performance among multilingual models (see https://www.sbert.net/docs/pretrained_models.html), and LaBSE because it has shown good performance on the task of aligning simplified and regular sentences (Stodden et al., 2023). The last model that we've selected is LASER (Artetxe and Schwenk, 2019), which is the default model for Vecalign. We are using the original LASER model because it has Finnish embeddings.

We use the NLTK (Bird et al., 2009) Punkt sentence tokenizer (Kiss and Strunk, 2006) for sentence segmentation in this work. Segmentation does not appear to be an issue for our data in most cases since it has been professionally proofread before publishing and then crawled from the original website as is.

4.1. Document alignment

We use two approaches to document alignment.

The first approach is, following the previous work (Dmitrieva and Konovalova, 2023), a simple comparison of document vectors made by averaging the embeddings of all sentences in a document. We use cosine similarity to find the closest vectors.

The second method that we use is a technique proposed in Thompson and Koehn (2020). First, we use the provided script[7] for obtaining document embeddings for candidate generation. This method can be used with different sentence embeddings, so we try it with all four types of embeddings mentioned above. We set the K nearest neighbors to 5 and keep all other parameters default (such as $J = 16$ and $\gamma = 20$). We also experiment with dimensionality reduction for all embeddings to see how different the results can be. Following the original paper (Thompson and Koehn, 2020), we set the new dimensionality to 128. For sentence-transformers, we use the dimensionality reduction technique proposed within the library[8]. For LASER, we use the PCA (principle component analysis) module from scikit-learn (Pedregosa et al., 2011).

Lastly, we use a simplified version of the candidate re-scoring method from Thompson and Koehn (2020) to re-score the output of the models that performed best during candidate generation. We only do this for the documents aligned with the Vecalign method and, following the original paper, use Vecalign with LASER sentence embeddings. Our formula for re-scoring is simply

$$S(E,F) = \frac{1}{len(E)} \sum_{e,f \in a(E,F)} sim(e,f) \qquad (1)$$

where E and F are the source and target documents respectively, $a(E,F)$ is the alignment between these documents, and sim is the cosine similarity between sentences. Unlike in the original paper, we do not divide by the total number of alignments, because the mismatch in sizes of source and target documents is so high that it does not make sense to penalize for unaligned sentences. Instead, we divide by the number of sentences in the Easy Finnish document, because that would be the maximum possible number of alignments. We also do not take into account the probability that both documents are in the correct language because our task is monolingual.

It should be noted that we treat document and sentence alignments as exclusive. So, if document 1 aligns with document 2, no other document can align with documents 1 or 2. In all document alignment methods, we employ a simple strategy to find the best match for each document after obtaining K best candidates. For all Easy Finnish documents, we find five possible Standard Finnish matches, obtaining a matrix of distances or similarities. Then, we find the maximum (for similarities) or minimum (for distances, which is what the Vecalign method returns) value in the matrix. We lock that document pair, eliminate it from the matrix, and look for the next highest or lowest value.

4.1.1. Evaluation

We use the Parallel Corpus of Finnish and Easy-to-read Finnish (Dmitrieva et al., 2022) to compare document alignment methods. This dataset has document pairs with "positive", "neutral", and "negative" labels. The "positive" label means that the human annotator working on the dataset was positive that the Easy Finnish document is the simplified version of the source document, the "negative" label means the opposite (the documents in the pair talk about different things), and the "neutral" label was given when the annotator was not sure. There are 1257 "positive", 470 "negative", and 192 "neutral" article pairs in the dataset (Dmitrieva and Konovalova, 2023). The labels were given after automatic pre-alignment had been performed, i.e. the annotator did not look for the pairs herself. During alignment evaluation, we only compare the document pairs that are present in both the predicted sample and the annotated dataset, so the support is different in every case. When counting the "strict" scores, we consider "neutral" documents to be positive, and when counting "lax" (relaxed) scores, we consider the "neutral" documents to be negative.

[7] https://github.com/thompsonb/vecalign/blob/master/standalone_document_embedding_demo.py

[8] https://github.com/UKPLab/sentence-transformers/blob/master/examples/training/distillation/dimensionality_reduction.py

We experimented with different thresholds for cosine similarity and distance scores. In our case, the distance is the cosine distance computed within the scikit-learn's nearest neighbors algorithm and defined as 1.0 minus the cosine similarity. For both metrics, there are 9 possible thresholds from 0.1 to 0.9. We have reached the conclusion that in the majority of cases, good F1 scores can be obtained with the highest (for distance) or lowest (for similarity) possible thresholds, which also let us obtain the highest number of pairs, i.e., have the best possible recall while still having high precision. Table 1 contains the evaluation results for the document alignment algorithm from Thompson and Koehn (2020) [the second approach], and Table 2 contains the results of document comparison with just cosine similarity between averaged sentence vectors [the first approach].

It appears that LaBSE and LASER embeddings are giving the best results in all cases. That is why we decided only to try the candidate re-scoring method (Thompson and Koehn, 2020) on the results obtained with these embedding models. However, in our case, candidate re-scoring proved not to be particularly helpful. Not only did the precision decrease, but we also got comparatively low support scores, which means that the set of document pairs that this algorithm retrieved matches the document pairs in the "true" data set rather vaguely. It can be seen that just the candidate generation algorithm from Vecalign worked best in our case. Using full-size embeddings as opposed to truncated embeddings gave only a slight improvement to the performance (same as in the original paper (Thompson and Koehn, 2020)), which means that in a more data-dense setting, truncated embeddings can be used.

4.2. Sentence alignment

For sentence alignment, we wanted the aligners to adhere to as many of the following criteria as possible:

- One-to-one, one-to-many, many-to-one, many-to-many sentence alignments are all possible.
- Crossing alignments/crossing links are allowed. Between document 1 with sentences A, B, C (here and in all examples below sentences are given in the exact order) and document 2 with sentences a, b, c, d, we can have alignments such as BC -> a and A -> d.
- Sentences within an alignment are consecutive. Between document 1 with sentences A, B, C and document 2 with sentences a, b, c, d, we cannot have alignments such as AC -> bd. We also cannot have alignments such as A -> ba; only A -> ab is possible.
- Alignments are exclusive. Between document 1 with sentences A, B, C and document 2 with sentences a, b, c, d, we cannot have both alignments A -> a and B -> a; only one of them can be chosen.
- If the method uses embeddings, it should be possible to change the embedding model.

We were unable to find a method that would satisfy all the criteria, so we opted for those that came close. We also designed a simple cosine similarity-based method to use as a baseline, satisfying all the above criteria. As another baseline, we use MASSAlign with TF-IDF-based text comparison, i.e. without any embeddings.

The first method that we use is **Vecalign** for sentence alignment (Thompson and Koehn, 2019). It is based on the similarity of sentence embeddings and a dynamic programming approximation, which is fast even for long documents. Vecalign is language-agnostic because it can work with any embeddings. It does not provide crossing alignments but satisfies all other requirements.

Our second aligner is **Bertalign** (Liu and Zhu, 2022), which works in two steps. The first step finds the optimal paths for 1-to-1 alignments based on the top-k most semantically similar target sentences for each source sentence using the bidirectional encoder representations from transformer-based cross-lingual word embeddings. The second step relies on search paths found in the previous step to recover all valid alignments with more than one sentence on each side of the bilingual text (ibid.). Bertalign outperforms Vecalign on the English-Chinese bilingual alignment (Liu and Zhu, 2022) and also on German-Easy German monolingual alignment (Stodden et al., 2023). This method also does not provide crossing alignments but satisfies all other requirements.

Both Vecalign and Bertalign let the user set the maximum number of consecutive sentences that can be aligned at once (maximum overlap size). We set this number to 3 in all experiments. We chose this threshold because in the manually aligned golden test set for sentence alignment evaluation described in paragraph 4.2.1, this is the maximum number of consecutive sentences appearing in one alignment, and 3:n and n:3 alignments are seen very rarely, so we did not see a reason to go over that limit.

We employ two baselines. The first is **MASSAlign** (Paetzold et al., 2017), which does not utilize embeddings at all. It uses a vicinity-driven approach in which it first creates a similarity matrix between the paragraphs/sentences of aligned documents/paragraphs, using a standard bag-of-words TF-IDF model, then finds a starting point to begin the search for an alignment path (ibid.). MASSAlign

		Strict			Lax				
Embeddings	**Dist↓**	**p**	**r**	**f1**	**p**	**r**	**f1**	**sup-1**	**sup-2**
Truncated embeddings									
LaBSE-128	0,9	0,723	**1,000**	0,840	0,820	**1,000**	**0,901**	1439	1439
MPNet-128	0,9	0,718	**1,000**	0,836	0,814	**1,000**	0,898	1453	1453
DistilUSE-128	0,9	0,712	**1,000**	0,832	0,808	**1,000**	0,894	1473	1473
LASER-128	0,9	**0,730**	0,993	**0,841**	**0,823**	0,993	0,900	1319	1329
Full-size embeddings									
LaBSE	0,9	0,728	**1,000**	0,842	0,824	**1,000**	0,903	1424	1424
MPNet	0,9	0,717	**1,000**	0,835	0,814	**1,000**	0,897	1473	1473
DistilUSE	0,9	0,711	**1,000**	0,831	0,807	**1,000**	0,893	1504	1504
LASER	0,9	**0,729**	**1,000**	**0,843**	**0,826**	**1,000**	**0,905**	1188	1188
After candidate rescoring									
LaBSE rescored	n/a	0,701	**1,000**	0,824	**0,805**	**1,000**	**0,892**	743	743
LASER rescored	n/a	**0,706**	**1,000**	**0,828**	0,803	**1,000**	0,891	595	595

Table 1: Document alignment with Vecalign document embeddings (Thompson and Koehn, 2020). "Sup-1" is support-1, the number of pairs deemed "positive" (true pairs) under the current threshold. "Sup-2" is support-2, the number of document pairs in the predicted sample that match the document pairs in the true dataset.

		Strict			Lax				
Embeddings	**Cos. sim.↑**	**p**	**r**	**f1**	**p**	**r**	**f1**	**sup-1**	**sup-2**
LaBSE	0,68	0,717	**1,000**	0,835	**0,812**	**1,000**	**0,896**	1613	1613
MPNet	0,55	0,701	**1,000**	0,825	0,797	**1,000**	0,887	1628	1628
DistilUSE	0,47	0,689	**1,000**	0,816	0,783	**1,000**	0,878	1710	1710
LASER	0,80	**0,719**	**1,000**	**0,836**	0,810	**1,000**	0,895	1574	1575

Table 2: Document alignment by comparing averaged sentence embeddings.

	Strict			Lax		
Embeddings	**p**	**r**	**f1**	**p**	**r**	**f1**
Vecalign						
LaBSE	0,786	0,305	0,439	0,847	0,7	0,766
MPNet	0,788	0,3	0,435	**0,852**	0,704	**0,771**
DistilUSE	0,789	0,314	0,449	0,841	0,65	0,733
LASER	**0,801**	**0,426**	**0,556**	0,839	0,668	0,744
Bertalign						
LaBSE	0,745	0,179	0,289	0,813	0,596	0,688
MPNet	0,77	0,269	0,399	0,822	0,601	0,694
DistilUSE	0,738	0,166	0,271	0,802	0,561	0,66
LASER	0,694	0,081	0,145	0,749	0,408	0,528
Cos. sim. matrix						
LaBSE	0,34	0,368	0,353	0,585	**0,726**	0,648
MPNet	0,304	0,305	0,304	0,607	0,691	0,646
DistilUSE	0,301	0,336	0,318	0,514	0,632	0,567
LASER	0,311	0,269	0,288	0,601	0,614	0,608
MASSAlign						
n\a	0,57	0,238	0,335	0,774	0,318	0,451

Table 3: Sentence alignment by different methods. "P" stands for "precision", "r" for recall, and "f1" for f1-score.

does not allow crossing alignments and sometimes returns non-exclusive alignments, but it has shown competitive results on the monolingual alignment task (Stodden et al., 2023; Spring et al., 2023). We use it with default values as in the example script[9], since we found out empirically that it is possible to obtain sensible alignments with these values. As a stop-words list, we use the stop-words list for Finnish from NLTK. The other baseline that we use is a simple algorithm similar to the one described in Section 4.1 for choosing the best documents out of K best. We embed all sentences and concatenations of consecutive sentences (of length 1 <= len <= 3) and obtain a **cosine similarity matrix**. Then, we look for the greatest value in this matrix, lock that alignment, eliminate all the sentences that go into that alignment (if we align sentences AB to sentence b, we must also eliminate rows A, B, ABC, BC, ab, abc, bc, bcd), and look for the next highest value. This method satisfies all our criteria.

4.2.1. Evaluation

We use the script provided in the Vecalign repository[10] to score our alignments. In order to obtain a gold test set, we manually aligned 50 randomly chosen "positive" document pairs from the Parallel Corpus of Finnish and Easy-to-read Finnish (Dmitrieva et al., 2022). There are 1638 singular sentences in Standard Finnish documents and 291 sentences in Easy Finnish documents. Between these documents, there are 223 non-zero alignments in the golden test set, of which 160 are one-to-one, 47 are one-to-many or many-to-one, and 16 are many-to-many ("many" was never higher than 3). The results can be viewed in Table 3.

It can be seen that Vecalign with LASER embeddings outperforms all other methods. Bertalign seems to work way worse on our data than, for example, on German monolingual data (Stodden et al., 2023). We have come to the conclusion that the performance of different alignment methods depends greatly on the nature of the data since even different monolingual corpora on the same language align differently: compare, for example, the results in Spring et al. (2023) and Stodden et al. (2023) that both deal with German-Easy German alignment. However, in Spring et al. (2023), Vecalign also demonstrated good performance. Unfortunately, we were unable to obtain good results with MASSAlign or Bertalign like Stodden et al. (2023) did. However, it should be noted that while annotating the golden test set, we concluded that a big part of our data may be difficult to align even for humans. The bigger the length difference between the Easy Finnish and Standard Finnish documents was, the harder it was to find true matches between the sentences.

	2019-20	**2014-18**	**Total**
Documents			
Pairs	1257	7004	8261
Words$_{reg}$	471565	1700469	2172034
Words$_{easy}$	69179	402274	471453
Sentences			
Pairs	2994	8950	11944
Words$_{reg}$	41056	116684	157740
Words$_{easy}$	26699	80926	107625

Table 4: Dataset statistics. "reg" stands for Standard Finnish, or regular, texts, "easy" stands for Easy Finnish. We only consider "positive" document pairs and sentence pairs with a score equal to or below 0.65.

Vecalign provides a score for all non-zero alignments, which reflects the cost of the alignment. The smaller the number is, the better the alignment. Zero scores are given to zero alignments (when the sentence is not aligned to any other sentence). We evaluated score thresholds from 0.1 to 0.9 on the golden test set and then empirically. To us, it appears that alignments with the score <= 0.65 can be confidently chosen for further use.

4.3. Dataset statistics

The statistics of our new dataset can be seen in Table 4. We have increased the amount of documents 6.5 times and added a sentence-aligned version of 11944 sentence pairs. We only considered pairs with the score <= 0.65. If the score limit is lifted, the total number of non-zero pairs in the entire dataset would be 56088.

5. Modeling

In addition to increasing the amount of Finnish simplification data, we also present the first baseline models for automatic Finnish sentence simplification. As mentioned before, we worked with two different architectures:

- mBART (Liu et al., 2020): a multilingual version of BART, a denoising autoencoder for pretraining sequence-to-sequence models, particularly effective when fine-tuned for text generation (Lewis et al., 2020). We use mBART cc25, a model with 12 encoder and decoder layers trained on 25 languages' monolingual corpus[11].

[9] https://ghpaetzold.github.io/massalign_docs/examples.html

[10] https://github.com/thompsonb/vecalign/blob/master/score.py

[11] https://github.com/facebookresearch/fairseq/tree/main/examples/mbart

	Highest SARI	Epoch
mBART	37.612	10
Finnish GPT	44.63	10

Table 5: Model evaluation results for sentence simplification.

Feature	mBART	FinnGPT	Target
Compression	0.710	0.680	0.743
Sentence splits	0.828	0.831	0.875
Levenshtein	0.782	0.610	0.559
Exact copies	0.181	0.036	0.020
Additions	0.057	0.297	0.403
Deletions	0.339	0.559	0.618

Table 6:
Quality estimation reports from EASSE.

- Finnish GPT: a Generative Pretrained Transformer with 1.5B parameters for Finnish. We use the XL version[12] and fine-tune it according to the authors' instructions[13].

We fine-tune mBART with default parameters as in the original instruction referenced above. For Finnish GPT, we employ instruction fine-tuning (Ouyang et al., 2022) and use the instruction "Mukauta selkosuomeksi" [translate to Easy Finnish]. We trained both models for 10 epochs.

For evaluation, we use the SARI metric, which uses an arithmetic average of n-gram precisions and recalls of editing operations: addition, keeping, and deletions between the source, output, and references (Xu et al., 2016). SARI is widely used for evaluating text simplification. It has some drawbacks, such as not being able to consider grammaticality or coherence, but it does have a good correlation with human judgments of simplicity (ibid.). Due to SARI's popularity, our results can be compared easily to any past works on simplification for other languages and future works on Finnish simplification. We use the code from the EASSE library (Alva-Manchego et al., 2019). The evaluation results can be seen in Table 5.

We also provide quality estimation features available in EASSE: the compression ratio of the simplification with respect to its source sentence, the Levenshtein similarity between source and simplification (calculated as Levenstein ratio in characters), the average number of sentence splits performed by the system, the proportion of exact matches (i.e. original sentences left untouched), the average proportion of added words and deleted words (Alva-Manchego et al., 2019). We do not report the lexical complexity score because, to the best of our knowledge, it is not language-agnostic in the current implementation. For comparison, we provide the quality estimation values between the source and target documents. The values can be seen in Table 6.

As can be seen, none of the systems has achieved the level of compression between the actual target sentences and source sentences. However, both mBART and Finnish GPT are close to the correct amount of sentence splitting. The higher Levenstein similarity, the number of exact copies, and the lesser amount of additions and deletions lead us to believe that mBART is a more conservative model, which can explain lower SARI scores.

Some examples of simplifications produced by models can be found in Table 7. It can be seen that mBART indeed makes fewer changes to the original sentence, sometimes leaving the smaller sentences unchanged. However, it should be noted that sometimes the target sentence also does not change the source much, so it is not necessarily an undesirable behavior. Finnish GPT seems to produce shorter and easier sentences but does not really change word order, which would be beneficial in the third example. The shortening can probably be explained by the fact that a lot of manual simplifications in our dataset also shorten the original sentences quite a lot. Sometimes, none of the models get to the simplification degree that the target sentence shows: for instance, in the 1st example, although the models performed some simplification, such as using a number to spell percentages and changing some words, it is still not as simple as the target. Nevertheless, the outputs are coherent and provide at least some degree of simplification.

The finetuned Finnish GPT can be found in the following repository: https://huggingface.co/Helsinki-NLP/simple-finnish-gpt3-xl. The finetuned mBART model can be accessed here: https://huggingface.co/Helsinki-NLP/fin-simple-mBART, or through the following PID: http://urn.fi/urn:nbn:fi:lb-2024011801.

6. Conclusions and future work

We have described the process of expanding a news-based parallel Finnish-Easy Finnish dataset, adding more document-aligned data and a sentence-aligned version. The augmented dataset spans almost 6.5 years of news broadcasts and has more than eight thousand document pairs and twelve thousand sentence pairs. We also made a small, manually aligned golden test set for sen-

[12] https://huggingface.co/TurkuNLP/gpt3-finnish-xl

[13] https://github.com/spyysalo/instruction-finetune

Source	Target	mBART	Finnish GPT
Suomen Apteekkariliiton mukaan tilastot osoittavat, että mielialalääkkeiden euromääräinen myynti on kasvanut Suomessa noin yhdeksän prosenttia viime vuodesta. According to the Finnish Association of Pharmacists, statistics show that the sales of mood drugs in euros in Finland have increased by around nine percent since last year.	Mielialalääkkeiden myynti on kasvanut Suomessa tänä vuonna. Sales of mood medications have increased in Finland this year.	Suomen Apteekkariliitto kertoo, että mielialalääkkeiden myynti on kasvanut Suomessa noin 9 prosenttia viime vuodesta. The Finnish Association of Pharmacists says that sales of mood medications have grown in Finland by about 9 percent since last year.	Suomen Apteekkariliitto sanoo, että mielialalääkkeiden myynti on kasvanut apteekeissa noin 9 prosenttia viime vuodesta. The Finnish Association of Pharmacists says that sales of mood medications have increased in pharmacies by about 9 percent since last year.
Onnettomuus tapahtui vähän ennen puolta päivää paikallista aikaa Coraton ja Andrian välisellä rataosuudella Etelä-Italiassa. The accident happened shortly before half a day at the local time on the line between Corato and Andria in southern Italy.	Onnettomuus sattui lähellä Andrian kaupunkia. The accident happened near the city of Andria.	Coraton ja Andrian välisellä rataosuudella Etelä-Italiassa on tapahtunut onnettomuus. An accident has occurred on the track section between Corato and Andria in southern Italy.	Onnettomuus tapahtui vähän ennen puolta päivää. The accident happened shortly before noon.
Sankarikoiria Suomessa on nimetty jo vuosia. Hero dogs in Finland have been named for years.	Suomessa on valittu sankarikoiria jo monta vuotta. Hero dogs have been chosen in Finland for many years.	Sankarikoiria Suomessa on nimetty jo vuosia. Hero dogs in Finland have been named for years.	Sankarikoiria on Suomessa. There are heroic dogs in Finland.

Table 7: Example simplifications. Finnish texts are from news articles (copyright: Yleisradio), and English texts are translations of the sentences above.

tence alignment. Currently, all these datasets can be found on Kielipankki (Dmitrieva and Yleisradio (2024a); Dmitrieva and Yleisradio (2024b)). We have obtained robust results on document alignment; however, despite trying multiple aligners that have been proven to work well for monolingual alignment, the predictive values for sentence alignment were not as high. Having worked on manual sentence alignment, we can conclude that it proves to be a genuinely difficult task to perform on our dataset. We leave a possible improvement of sentence alignment for future work. Nevertheless, sentence simplification models perform fine on our sentence-aligned data in comparison to SARI scores obtained on other languages (see, for example, the fine-tuning experiment results in Ryan et al., 2023). We hope that our results can be used as a baseline for future works on Finnish sentence simplification. Another prospective task that we see is document-level simplification for Finnish. Having a good-quality document-aligned dataset will allow for experimenting with full document simplification and/or document-level planning for simplification (Cripwell et al., 2023).

7. Ethical considerations and limitations

The data described in this research is available on Kielipankki for non-commercial use. Datasets based on texts from the Yle archives cannot be deposited elsewhere for copyright reasons. Only people with login credentials from certain academic organizations or those who have obtained permission from Kielipankki will be able to download this data.

We cannot guarantee that all automatically aligned sentence or document pairs are correctly aligned. As mentioned above, due to the difficult nature of sentence alignment across our data, some erroneous sentence alignments can be expected even when the score threshold is in place. We kept the cost scores provided by Vecalign in the published data for transparency.

We acknowledge that text simplification models' output cannot be thoroughly evaluated with just automatic metrics because they do not assess grammaticality or coherence. However, we hope that increasing the amount of available simplification data will help the development of more sophisticated data-driven simplification evaluation approaches, such as LENS (Maddela et al., 2023), for languages other than English.

Most computations that required GPU, which are embedding operations and model training, were performed with a single GPU node, the GPU being a Nvidia Tesla V100 with an Xeon Gold 6230 processor. Running Finnish GPT fine-tuning with LoRA (Hu et al., 2021) required two nodes with 48 gigabytes of memory allocated per node, although we are unsure if this is the minimum memory requirement (i.e., the minimum requirement might be smaller).

8. Bibliographical References

Rodrigo Alarcón, Paloma Martínez, and Lourdes Moreno. 2023. Tuning BART models to simplify Spanish health-related content. *Procesamiento del Lenguaje Natural*, 70:111–122.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. 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): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a parallel Finnish-Easy Finnish dataset from news articles. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 21–26, Tampere, Finland. European Association for Machine Translation.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic text simplification for German. *Frontiers in Communication*, 7:706718.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

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, Tilburg, the Netherlands. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Auli Kulkki-Nieminen. 2010. *Selkoistettu uutinen. Lingvistinen analyysi selkotekstin erityispiirteistä [Plain Language news: a linguistic analysis of the sfecial features of simplified text]*. Ph.D. thesis, Tampereen Yliopisto.

Leealaura Leskelä. 2021. Easy language in Finland. In Ulla Vanhatalo Camilla Lindholm, editor, *Handbook of Easy Languages in Europe*, 1 edition, volume 8 of *Easy – Plain – Accessible*, pages 149–190. Frank & Timme.

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.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

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.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4, Tapei, Taiwan. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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, Online and in Dominican Republic. Association for Computational Linguistics.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

A.A. Shatilov and A.I. Rey. 2021. Sentence simplification with ruGPT3. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 1–13.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Nicolas Spring, Marek Kostrzewa, David Fröhlich, Annette Rios, Dominik Pfütze, Alessia Battisti, and Sarah Ebling. 2023. Analyzing sentence alignment for automatic simplification of German texts. In Silvana Deilen, Silvia Hansen-Schirra, Sergio Hernández Garrido, Christiane Maaß, and Anke Tardel, editors, *Emerging Fields in Easy Language and Accessible Communication Research*, pages 339–369. Frank & Timme GmbH, Berlin.

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 International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Hideki Tanaka, Hideya Mino, Tadashi Kumano, Shinji Ochi, and Motoya Shibata. 2013. News service in simplified Japanese and its production support systems. *The Best of IET and IBC*, 5:44–48.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

9. Language Resource References

Anna Dmitrieva and Aleksandra Konovalova and Yleisradio. 2022. *Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2019-2020, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2022111625.

Anna Dmitrieva and Yleisradio. 2024a. *Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2014-2018, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2024011701.

Anna Dmitrieva and Yleisradio. 2024b. *Parallel Sentence Aligned Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2014-2020, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2024011703.

Yleisradio. 2017. *Yle Finnish News Archive 2011-2018*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2017070501.

Yleisradio. 2019. *Yle News Archive Easy-to-read Finnish 2011-2018, source*. Kielipankki. PID http://urn.fi/urn:nbn:fi:lb-2019050901.