Creating a parallel Finnish–Easy Finnish dataset from news articles

Anna Dmitrieva
University of Helsinki
Yliopistonkatu 4, 00100 Helsinki, Finland
anna.dmitrieva@helsinki.fi

Aleksandra Konovalova
University of Turku
FI-20014 Turun yliopisto, Finland
aleksandra.a.konovalova@utu.fi

Abstract

Modern natural language processing tasks such as text simplification or summarization are typically formulated as monolingual machine translation tasks. This requires appropriate datasets to train, tune, and evaluate the models. This paper describes the creation of a parallel Finnish–Easy Finnish dataset from the Yle News archives. The dataset contains 1919 manually verified pairs of articles, each containing an article in Easy Finnish (*selkosuomi*) and a corresponding article from Standard Finnish news. Standard Finnish texts total 687555 words, and Easy Finnish texts have 106733 words. This new aligned resource was created automatically based on the Yle News archives from the Language Bank of Finland (Kielipankki) and manually checked by a human expert. The dataset is available for download from Kielipankki. This resource will allow for more effective Easy Language research and for creating applications for automatic simplification and/or summarization of Finnish texts.

1 Introduction

Easy and Plain Languages can be considered language varieties of different national languages with reduced linguistic complexity, which aim to improve the readability and comprehensibility of texts (Maaß, 2020). However, "Easy Language" is not exactly a uniform concept, but more of an umbrella term (Lindholm and Vanhatalo, 2021). Easy Language media in the Nordic countries have a long history, with the first documented signs of explicit usage of Easy Language in Sweden dating back to the 1960s (ibid.). In Finland, the first books and magazines in Easy Finnish were published in the early 1980s (Leskelä, 2021). Right now, Easy Language is well-established in Finland in practice, and the general attitude towards it is mainly positive (ibid.). Archives of *selkosuomi* [Easy Finnish] texts are available in the Language Bank of Finland (Kielipankki[1]). However, despite the availability of these resources, there seemingly has been no effort to create a sizeable aligned parallel corpus. Our research aims to fill this gap.

Our corpus is based on the Yle news archives available on Kielipankki (Yle Finnish News Archive 2019–2020[2] and Yle News Archive Easy-to-read Finnish 2019–2020[3]). Yle is a Finnish public service media company. It provides content for different audiences, including special target groups such as Easy Finnish language users. *Yle Uutiset selkosuomeksi* [Yle News in Easy Finnish] provides short (about 5 minutes) daily radio and TV broadcasts. The radio broadcasts are also published on Yle's website in text form. In this paper, our focus is on the news in text format.

The reason for creating this corpus is twofold. First, we wanted to provide a resource for studying the simplification strategies used by simple language content providers. This task requires parallel data in which Standard Finnish texts are aligned with their simplified versions. Second, we wanted to start working towards creating automatic text simplification tools for Finnish. Nowadays, automatic text simplification is typically seen

[1] https://www.kielipankki.fi/
[2] http://urn.fi/urn:nbn:fi:lb-2021050401
[3] http://urn.fi/urn:nbn:fi:lb-2021050701

Esplà-Gomis, Forcada, Kuzman, Ljubešić, van Noord, Ramírez-Sánchez, Tiedemann, Toral (eds.)
Proceedings of the 1st Workshop on Open Community-Driven Machine Translation, p. 21–26
Tampere, Finland, June 2023.

as a monolingual machine translation (MT) problem (Kriz et al., 2019), and MT models require parallel text data to train on. In this stage of our research, we have developed and applied the methodology for text alignment and manually assessed the alignments. The current dataset is available online and can be used for training models in low-resource settings. The Standard Finnish articles have 687555 words in total, and the Easy Finnish articles amount to 106733 words.

2 Related work

Parallel datasets for monolingual machine translation are often created based on news. For example, the CNN / DailyMail dataset (Hermann et al., 2015) of unique news articles written by journalists at CNN and the Daily Mail is used for abstractive and extractive summarization, and the ParaPhraserPlus corpus (Gudkov et al., 2020), which contains pairs of similar news headlines, is used for paraphrase generation. News-based corpora are also used for simplification, such as the Newsela corpus, which consists of news articles simplified manually by professional editors (Xu et al., 2015).

To align Easy Finnish articles with Standard Finnish equivalents, we used sentence embeddings. Multilingual sentence embeddings are often applied in paraphrase identification: for example, Girrbach (2022) showed that publicly available, pretrained models already achieve solid performance on this task. Khairova et al. (2022) also prove the advantages of using the fine-tuned Sentence-BERT language model for classifying paraphrased sentences over other modern models.

Current research on Easy Finnish covers many different resources and methods. One of the most recent papers containing corpus-based research focuses on the meanings of the word *ihminen* [human] and its usage in Easy Finnish and Standard Finnish news articles (Valtasalmi, 2021). It is a good example of research that could potentially benefit from our parallel corpus. Leskelä (2022) takes spoken Easy Finnish as a primary topic, and Hyppönen (2022) focuses on the cognitive accessibility assessment of different service websites.

According to the interviews conducted by Kulkki-Nieminen (2010), the main three target groups for news in Easy Finnish are immigrants, older adults, and people with intellectual disabilities. When *Yle Uutiset selkosuomeksi* started broadcasting, however, it was primarily aimed at heritage Finnish speakers who had learned Finnish at home but were no longer in the Finnish language environment (P. Seppä, personal communication, March 3, 2023). Kulkki-Nieminen (2010) conducted the interviews about the main topic of their research, namely the linguistic analysis of the language feature specific to Easy Finnish texts.

3 Dataset creation

3.1 Data selection and preprocessing

Kielipankki has multiple archives of Yle news. When we started creating our dataset, there was a more extensive archive comprising articles from 2011–2018 and a smaller archive for 2019–2020. To test our methodology and, at the same time, work with more modern texts, we chose the smaller archive.

Yle News in Easy Finnish and general Yle news, as well as the Swedish Yle partition, are stored in different archives. These archives are in JSON format and contain all available information about each article, including image captions within the text, topics, subjects, etc. Upon looking at the regular Finnish news archive, we noticed that Easy Finnish news articles are often mixed into it, so we had to filter out such occasions, removing entries with a "selkouutiset" topic. We did not perform additional preprocessing on the articles apart from cleaning out noise (non-Latin symbols, emojis, etc.).

3.2 Reducing the search space

There is no explicit alignment between the Easy Finnish news. Before looking for pairs of texts, we had to decide to only look for pairs between articles that came out on the same day. As we learned later, it is a valid approach since most regular news articles, if selected for Easy Finnish news, are translated and come on air within 24 hours (P. Seppä, personal communication, March 3, 2023). However, in some cases, non-urgent matters are covered in Easy Finnish news later.

The Easy Finnish news from each day's radio broadcast is usually combined on one page, and each paragraph covers its own event. Therefore, we matched these paragraphs to Standard Finnish news from the same day. Since we were looking for Standard Finnish equivalents of *selkosuomi* news in a one-to-all fashion, meaning that we looked through all Standard Finnish articles each time in order to find a match for a single *selkosuomi* article, we wanted to limit the number of Standard Finnish articles we have to look through at each step. Therefore, we did not just limit matching to articles from the same day but also only looked at Standard Finnish articles with some of the same subjects that the *selkosuomi* articles from this day had. Sometimes the articles are translated into Easy Finnish from Swedish (P. Seppä, personal communication, March 3, 2023), but for this project, we limited ourselves to matching only Finnish articles.

3.3 Alignment strategies

After the article matching explained in the previous section, we tried different techniques to find pairs of Finnish and Easy Finnish articles discussing the same topic.

The first strategy we tested was to try and match some articles based on the same image captions. For that, we lifted the same-day limitation. Unfortunately, upon inspection, this strategy proved to be unreliable, so we did not use it. Some captions were too broad to guarantee that all articles in which the corresponding image is found are on the same topic. However, we do not completely abandon this strategy and acknowledge that paired with paraphrase identification techniques, it could be used for expanding the dataset.

In order to find pairs of equivalent articles, we needed a tool to estimate the similarity between two texts. To do so, we looked at Doc2Vec and Sentence Transformers to obtain document embeddings that can be used to measure the cosine similarity between them.

Doc2Vec is a model that can represent a document (i.e., a text of one or more sentences) as a single vector. We used the Gensim implementation[4] based on Le and Mikolov (2014) and a SentencePiece[5] unigram language model (Kudo, 2018) for tokenization. To train the SentencePiece and Doc2Vec models, we used the Yle News archives from 2016–2018,[6] something not yet included in our dataset. A total of 202,656 articles were used for Doc2Vec, and 1 million randomly selected sentences for SentencePiece.

Sentence Transformers are language models used to derive semantically meaningful sentence embeddings. For this project, we used one of the models from SBERT[7] (Reimers and Gurevych, 2019): the multilingual knowledge-distilled version of multilingual Universal Sentence Encoder (Yang et al., 2020), version 2 (distiluse-base-multilingual-cased-v2). Version 1 is said to have better performance but does not include Finnish, so we opted for the second version. In order to get an embedding of a document, we take the average of all sentence vectors in the document. Since this model makes a vector for each individual sentence, transforming larger articles into vectors can be computationally expensive, so we only use the first 15 sentences to make an article's vector. It should be noted that the length of most Easy Finnish articles does not exceed this limit.

We manually compared both approaches on a set of Standard Finnish and Easy Finnish articles from the same random date. After looking at a sample of Standard and Easy Finnish articles from a few days, we found out that, even though the Sentence Transformer operated with only the first 15 sentences of the documents, as opposed to Doc2Vec, which utilized the entire documents, the Sentence Transformer performed better in finding equivalent articles. It also gave more representative scores: true pairs received high similarity scores (> 0.6), and false pairs received low scores (≤ 0.3), as opposed to the Doc2Vec model, where the scores were between 0.47 and 0.6. It is possible that a larger Doc2Vec model could have given better results. Still, for the sake of convenience and reproducibility of our work, we chose to proceed with the Sentence Transformer architecture for measuring semantic similarity.

3.4 Similarity threshold

For each Easy Finnish article in our collection, we found a Standard Finnish pair by choosing the article with the highest cosine similarity. Sometimes, it was impossible to find a good match. We had to establish a threshold of cosine similarity because, obviously, lower scores indicate a higher possibility of a false match. As seen in Figure 1, most pairs have cosine similarity scores between 0.6 and 0.7, with a little less than 500 pairs having a score of 0.5. We decided to perform the human evaluation on articles with scores from 0.6 to 1 to assess whether the scores actually represent semantic similarity and, if so, what the threshold should be.

[4] `https://radimrehurek.com/gensim/models/doc2vec.html`

[5] `https://github.com/google/sentencepiece`

[6] `http://urn.fi/urn:nbn:fi:lb-2017070501`

[7] `https://www.sbert.net/`

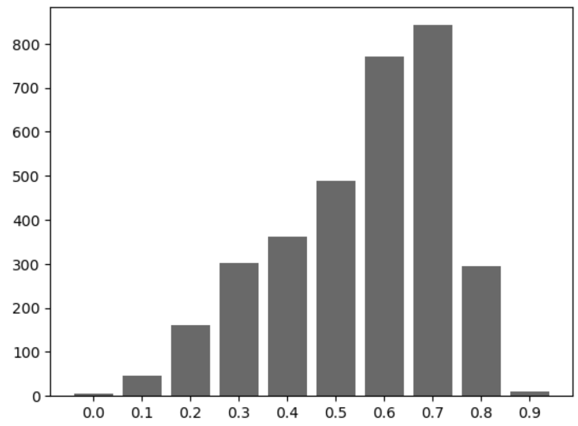

Figure 1: Distribution of cosine similarity scores across article pairs. X-axis is the approximated cosine similarity, y-axis is the number of pairs.

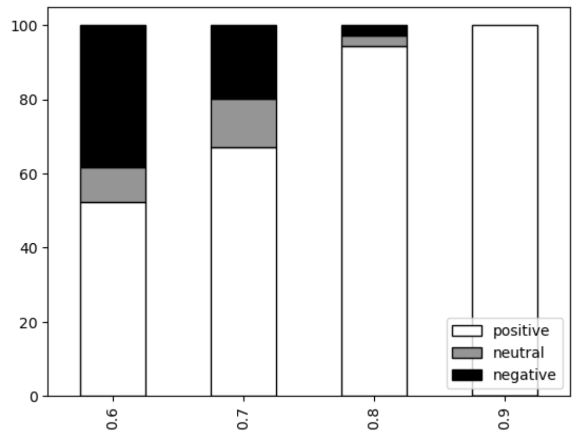

Figure 2: Percentages of labels given by the expert. X-axis is the approximated cosine similarity, y-axis is the percentage.

4 Evaluation

We ended up with 1919 pairs of articles with high (≥ 0.6) similarity scores. An expert was asked to evaluate each pair and give it one of three scores: "positive" – if the articles are definitely about the same topic, "negative" – if the articles definitely talk about different topics, or "neutral" – if it cannot be definitively said whether or not the articles talk about the same topic. The expert also gave comments on most of the negative cases. As seen in Figure 2, the percentage of negative labels grows with the decline of the similarity score. Therefore, we conclude that the cosine similarity indeed represents the semantic similarity of texts as seen by a human expert, and since the percentage of positive texts with labels between 0.6 and 0.7 is approx. 52%, there is no need to decrease the lower similarity threshold. There are 1257 "positive", 470 "negative", and 192 "neutral" article pairs in the dataset. Therefore, 65.5% of the data is "positive", 24.5% is "negative", and 10% is "neutral". These assessments can be used for classifying article pairs automatically, allowing for the creation of larger parallel datasets from the remaining Yle news archives.

The most common reasons for giving a pair of articles a "negative" or "neutral" score were as such:

- Easy Finnish article is about a completely different topic.
- Easy Finnish article covers a similar topic but does not exactly match the original article for various reasons (e.g., time, location, different focus).
- Easy Finnish article cannot be mapped to one original article but compiles information from several original articles.

It should be noted that some topics like Brexit, coronavirus, or weather forecasts were challenging to evaluate, especially when no particular time markers were mentioned. Due to the number of news on the same topic being too high, it was sometimes difficult to establish if two similar articles were definitely talking about the same event. There were also cases when the Easy Finnish article covered a topic relevant to the entire country of Finland, but the Standard Finnish article was limited only to one particular region (e.g., Lapland).

During the assessment, the expert compared the simplification strategies they saw in the Easy Finnish articles to the Easy Finnish Indicator/Selkomittari 2.0.[8] Selkomittari is a document that outlines the criteria for assessing how simple the text is. However, we did not use it to determine the quality of simplification but to see which strategies Easy Language content authors use in their work. According to the expert, the following characteristics are present in most of the simplified articles:

- The text mainly contains general vocabulary evaluated as familiar to the readers.
- The text does not contain lots of long words.
- The text contains no figures of speech that require creative reasoning to understand (to chip away at something, brain drain, etc.).

[8] https://selkokeskus.fi/wp-content/uploads/2022/04/Selkokielen-mittari-2.0.pdf

index in selko	index in regular	selko text	regular text	cos_sim	status	comments
3-10973979_0	3-10972641	Raakaöljyn hinta on noussut tänään melkein 10 prosenttia. Hinnannousun syy ovat Saudi-Arabiaan lauantaina tehdyt iskut...	Öljyn hinta nousi enemmän kuin Iranin vallankumouksen tai Kuwaitin sodan alettua. Öljyn hinta lähti odotetusti jyrkkään nousuun markkinoiden avauduttua maanantaiaamuna...	0,84402	Positive	Small difference in selkonews: last sentence

Table 1: An example of a dataset entry. The "index in selko" column includes the index of the entire entry in the Kielipankki dataset and the paragraph number after the underscore. Copyright: Yleisradio Oy, Finnish Broadcasting Company (Yle).

- The text contains high, precise numerical figures only if this is justified by its topic. If necessary, figures are approximated.
- Figures, numbers, units of measure, and relationships between numbers are presented visually.
- The text contains no abbreviations or acronyms, except for established ones which are better recognized as abbreviations than if written out in full (PDF, DVD).
- The text does not contain many language structures rated difficult.
- Clauses and sentences are mainly short.
- The text contains no words that have several different elements, such as derivative affixes, inflectional suffixes, and clitics.
- Sentence structures are simple. For the most part, they only have one subordinate clause.

It should be noted that in some cases, the vocabulary of Easy Finnish articles was not easy. For example, the expert has encountered the word *amurinleopardikissapariskunta* (the pair of Amur leopards), which was not used even in the original news article. Other cases of not-so-easy linguistic constructions found by the expert include complex sentence structures (in one case, a long sentence with four subordinate clauses) and colloquialisms without any additional comments on their meaning. There were also Easy Finnish articles that just summarized the original ones with no sentence-level simplification.

Obviously, not everything needs to be simplified. Sometimes, the authors of Easy Finnish texts will use complex words or constructions if they consider it necessary based on their expertise. The only reason we need to point out that sometimes the Easy Finnish articles might have "difficult" sentences is for other researchers to be aware of such cases. For example, suppose someone wants to make a sentence-aligned dataset based on Standard and Easy Finnish news to train an automatic text simplification model on. In that case, filtering out the not-so-easy Easy Finnish sentences may be beneficial before training.

5 Conclusions

We have described the creation of a parallel Finnish–Easy Finnish dataset based on news articles. With the help of a human assessor, we evaluated the automatically aligned pairs of articles, which helped us to determine the optimal similarity threshold and identify various cases of incorrect or ambiguous alignments. We also describe some of the simplification strategies used by authors during the creation of Easy Finnish articles.

This resource is now available for download on Kielipankki under the CLARIN ACA – NC license: `http://urn.fi/urn:nbn:fi:lb-2022111625`. The dataset can be used to study the linguistic properties of simplified Finnish and for various natural language processing applications. For example, it can be used for low-resource simplification or summarization. An example dataset entry can be seen in Table 1.

Some steps can be taken to improve our data collection procedures further. For example, a valuable contribution would be to find a way to match Easy Finnish articles to Swedish sources or Standard Finnish ones from different dates without it being too computationally expensive. Currently, we are working on extending the dataset and adding sentence-to-sentence alignments to create a more extensive dataset suitable for text simplification.

Acknowledgements: Anna Dmitrieva's work on this paper was partially supported by the Finnish Cultural Foundation (Grant №00230286).

References

Girrbach, Leander. 2022. PAR-MEX shared task submission description: Identifying Spanish paraphrases using pretrained models and translations. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Gudkov, Vadim, Olga Mitrofanova, and Elizaveta Filippskikh. 2020. Automatically ranked Russian paraphrase corpus for text generation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 54–59, Online, July. Association for Computational Linguistics.

Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Cortes, C., N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Hyppönen, Annikki. 2022. ”Hyvä saavutettavuus hyödyttää kaikkia” - Kognitiivisen saavutettavuusarvioinnin käytänteitä [Good accessibility benefits everyone: cognitive accessibility assessment practices]. In H. Katajamäki, M. Enell-Nilsson, H. Kauppinen-Räisänen H. Limatius, editor, *Responsible Communication*, volume 14, pages 43–59.

Khairova, Nina, Anastasiia Shapovalova, Orken Mamyrbayev, Nataliia Sharonova, and Kuralay Mukhsina. 2022. Using BERT model to identify sentences paraphrase in the news corpus. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022). Volume I: Main Conference*, volume 3171, pages 38–48.

Kriz, Reno, Joao 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 NAACL-HLT*, pages 3137–3147.

Kudo, Taku. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Kulkki-Nieminen, Auli. 2010. *Selkoistettu uutinen. Lingvistinen analyysi selkotekstin erityispiirteistä [Simplified news article. Linguistic analysis of special features of Easy Language text]*. Ph.D. thesis.

Le, Quoc and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In Xing, Eric P. and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun. PMLR.

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

Leskelä, Leealaura. 2022. *Selkopuhetta!: Puhuttu selkokieli kehitysvammaisten henkilöiden ja ammattilaisten vuorovaikutuksessa [Speak Easy Language! Spoken Easy Language in interactions between persons with intellectual disabilities and professionals]*. Ph.D. thesis.

Lindholm, Camilla and Ulla Vanhatalo. 2021. Introduction. In Camilla Lindholm, Ulla Vanhatalo, editor, *Handbook of Easy Languages in Europe*, volume 8 of *Easy – Plain – Accessible*, pages 11–26. Frank & Timme, 1 edition.

Maaß, Christiane. 2020. *Easy Language – Plain Language – Easy Language Plus*, volume 3 of *Easy – Plain – Accessible*. Frank & Timme, 1 edition.

Reimers, Nils 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Valtasalmi, Idastiina. 2021. Selkoa ihmisestä: Ihminen-sanan merkitykset ja käyttö selkokielisissä sanomalehtiteksteissä [Easy about a human: Human-word meanings and usage in Easy Language news articles]. *Sananjalka*, 63, March.

Xu, Wei, 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.

Yang, Yinfei, 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, July. Association for Computational Linguistics.