Proceedings of the 2nd READI Workshop @ LREC2022, pages 25–30
Marseille, 24 June 2022

A Multilingual Simplified Language News Corpus

Renate Hauser, Jannis Vamvas, Sarah Ebling, Martin Volk
Department of Computational Linguistics, University of Zurich
renateines.hauser@uzh.ch, {vamvas, ebling, volk}cl.uzh.ch

Abstract

Simplified language news articles are being offered by specialized web portals in several countries. The thousands of articles that have been published over the years are a valuable resource for natural language processing, especially for efforts towards automatic text simplification. In this paper, we present SNIML, a large multilingual corpus of news in simplified language. The corpus contains 13k simplified news articles written in one of six languages: Finnish, French, Italian, Swedish, English, and German. All articles are shared under open licenses that permit academic use. The level of text simplification varies depending on the news portal. We believe that even though SNIML is not a parallel corpus, it can be useful as a complement to the more homogeneous but often smaller corpora of news in the simplified variety of one language that are currently in use.

Keywords: Corpus, Multilinguality, Simplified Language, Accessibility, Text Simplification

1. Introduction

Simplified languages[1] are language varieties that have the purpose of enabling inclusion and social participation of people with low reading competence by making information easier to read and comprehend (Bredel and Maaß, 2016). A typical application domain of simplified language is news articles (Saggion, 2017). In recent years, web portals for simplified news have been created in several languages, and computational approaches to simplified language such as automatic readability assessment or text simplification are gaining interest.

We present SNIML, a corpus of **s**imple **n**ews **i**n **m**any **l**anguages. The corpus contains simplified news articles in Finnish, French, Italian, Swedish, English, and German. It comprises a total of 13,400 articles published between 2003 and 2022. All texts in SNIML are shared under an open license that allows for academic research use. We plan that future news articles are automatically collected and added to the corpus in a temporally stratified way. The corpus and various subcorpora are available for download.[2]

While some resources for simplified language align simplified versions of texts to standard-language versions, our corpus currently does not contain any news in standard language. Furthermore, the sub-corpora involve texts created according to different simplification guidelines and for different target audiences (Section 3.3). As such, SNIML is particularly useful for automatic readability assessment and for unsupervised, self-supervised or cross-lingual learning. In addition, the simple news articles could be combined with related news articles in standard language, yielding a large-scale multilingual comparable corpus of simple news.

Standard language

One possible difference is that Omicron may be less likely than earlier variants to cause a loss of taste and smell. Research suggests that 48 percent of patients with the original SARS-CoV-2 strain reported loss of smell and 41 percent reported loss of taste, but an analysis of a small Omicron outbreak among vaccinated people in Norway found that only 23 percent of patients reported loss of taste, and only 12 percent reported loss of smell.

Simplified language

Omicron may be less likely to cause a loss of taste and smell. The original virus caused such losses in almost half the sick people, and a study showed less than half that number losing taste and smell with Omicron.

Figure 1: Examples of news text in standard language and simplified language.

2. Related Work

Many previous resources for the computational processing of simplified language are *parallel*, combining a simplified version of a text with its original version. Parallel corpora have been used to train sequence-to-sequence systems for text simplification (Wubben et al. (2012); Nisioi et al. (2017); among others).

Notable examples of parallel corpora include the PorSimples corpus of Brazilian Portuguese news and popular science articles (Aluísio and Gasperin, 2010), the Simplext corpus of Spanish news (Bott and Saggion, 2014), the Newsela corpus of English and Spanish news (Xu et al., 2015), the Alector corpus of French educational texts (Gala et al., 2020), as well as German parallel corpora compiled from various web

[1] The term "simplified language" is used to denote the sum of all "comprehensibility-enhanced varieties of natural languages" (Maaß, 2020, p. 52), i.e., what is commonly termed "Easy Language" (German *leichte Sprache*) and "Plain Language" (German *einfache Sprache*). Maaß (2020, p. 52) mentions "easy-to-understand language" as an umbrella term subsuming these varieties. However, in this contribution, we prefer the term "simplified language" to emphasize the notion of the result of a simplification process.

[2] https://pub.cl.uzh.ch/projects/sniml/

sources (Klaper et al., 2013; Battisti et al., 2020). Other parallel news datasets that have been used for automatic simplification in German are the APA dataset (Säuberli et al., 2020) and the 20m dataset (Rios et al., 2021).

Some other previous resources for simplified language are *comparable* corpora, i.e, collections of simplified documents and standard-language documents that share the same topic but are not guaranteed to correlate on the sentence level. For example, Barzilay and Elhadad (2003) combined Encyclopedia Britannica and Britannica Elementary to form a comparable corpus, and they proposed to use alignment techniques to extract parallel sentence pairs. Similarly, the combination of English Wikipedia and Simple English Wikipedia (Zhu et al., 2010), though often treated as a parallel corpus, can be characterized as a comparable corpus, since Simple English entries are not necessarily simplified versions of the standard-language entries. Hwang et al. (2015) used parallel corpus mining methods to create a more parallel version of this dataset. An alternative approach is to extract parallel sentences manually from comparable corpora (Grabar and Cardon, 2018).

The SNIML corpus differs from previous resources with regard to its scale and its rich multilinguality. However, it is neither a parallel nor a comparable corpus, since it currently does not contain standard-language news articles. Thus, the structure of the corpus is best compared to datasets of raw text, such as the CC-News crawl (Common Crawl, 2016) or the Oscar corpus (Centre Inria de Paris, Équipe ALMAnaCH, 2019), even though SNIML is much smaller in size.

3. Corpus

3.1. Data

Language Variety The corpus is a collection of news articles that are written in simplified language. The articles originate from news providers in the USA, Finland (Finnish and Swedish), Belgium (French), Italy, and Switzerland (German). The level of simplification varies between the providers. Also, diverse target groups are addressed by the articles, including people with intellectual disabilities, people with low education, immigrants, emigrants, language learners in general, older adults, and children. The news providers are described in more detail in Section 3.2. Section 3.3 discusses the simplification guidelines involved.

Dataset Statistics Table 1 provides statistics of the corpus and its sub-corpora. The corpus consists of 13,447 articles that were published by the news providers listed below. The lengths of the articles vary greatly, within one provider as well as among the different platforms. *Journal Essentiel* and *Infoeasy* tend to publish longer articles, averaging above 600 tokens per article, while the average for the other providers generally lies below 300 tokens per article.

Temporal Stratification It is planned that news articles continue to be automatically fetched from the web and are added to new versions of the corpus. We plan to release a new version of SNIML every month.

Language	Articles	Sentences	Tokens
Finnish	3,379	41,792	661,194
French (BE)	2,723	102,496	1,759,518
Italian (IT)	2,686	10,824	737,903
Swedish (SV)	2,559	32,145	621,879
English (US)	1,897	50,999	965,805
German (CH)	147	25,964	123,021
Total	13,447	268,350	4,936,181

Table 1: Corpus statistics.

Machine Translations We have created English and German machine translations of most articles. These are mainly intended to be an aid for the users of the web reader interface (Section 4). In addition, the translations could be useful for data augmentation. We provide translations only in cases where the license permits derivative work, namely for articles provided by *Informazione Facile*, *Journal Essentiel* and *The Times in Plain English*.[3]

3.2. News Providers

We collected the articles of six news providers: *Selkosanomat*, *Lätta Bladet*, *Journal Essentiel*, *Informazione Facile*, *The Times in Plain English* and *Infoeasy*. Table 2 gives an overview of the providers and the licenses these providers apply to the text content. Text samples for each provider are listed in the Appendix (Table 3). In what follows, each provider is described in more detail.

3.2.1. Selkosanomat

Selkosanomat is a news platform in Finland that is published by the association Selkokeskus which is part of Kehitysvammaliitto (Developmental Disability Association). It offers a printed magazine as well as a free online newspaper. News on the topics Finland, world, sports, culture, and everyday life are published. The articles are written in Finnish.

3.2.2. Journal Essentiel

Journal Essentiel is an online journal in Belgium that is published by the non-profit association FUNOC (Formation pour l'Unversité ouverte de Charleroi) and primarily aims to be an educational information tool. The articles address current topics and are written in French.

[3]We used the commercial machine translation system of Microsoft Azure Cognitive Services (version 3.0). A manual investigation of the MT quality revealed typical machine translation errors such as wrong pronomina or incorrect gender. While we did not spot translation problems specific to simplified language, future work could investigate this question in more depth.

Provider	URL	Language	License
Selkosanomat	https://selkosanomat.fi/	fi	CC BY-NC-ND 4.0
Journal Essentiel	https://journalessentiel.be/	fr-BE	CC BY-SA 4.0
Informazione Facile	https://informazionefacile.it/	it-IT	CC BY-SA 4.0
Lätta Bladet	https://ll-bladet.fi/	sv-SE	CC BY-NC-ND 4.0
The Times in Plain English	https://www.thetimesinplainenglish.com/	en-US	*"may be reproduced and distributed by all"*
Infoeasy	https://infoeasy-news.ch/	de-CH	CC BY-NC-ND 4.0

Table 2: List of providers of the news articles that constitute the corpus.

3.2.3. Informazione Facile

Informazione Facile is an Italian online news platform published by the non-profit association *IF Informazione Facile*. Many topics are covered by the platform, including international news, Italian news, society and culture, sports, and health.

3.2.4. Lätta Bladet

Lätta Bladet is a sister magazine of *Selkosanomat* that is also situated in Finland and offers articles in Swedish. It is published by Selkokeskus and LL-Center, which is part of the interest organization of Swedish-speaking people with intellectual disabilities in Finland FDUV. Parallel to *Selkosanomat*, a printed magazine as well as a free online newspaper is offered. The same topics are covered by *Lätta Bladet* as by *Selkosanomat*. The articles are written in Swedish.

3.2.5. The Times in Plain English

The Times in Plain English is located in the USA and is published by the *News in Plain English Inc*. A wide range of topics are covered, including international news, news about New York, politics, health and education, law, and economy. To test readability, the publishers use Flesch-Kincaid Grade level (Kincaid et al., 1975).

3.2.6. Infoeasy

Infoeasy is a private initiative in Switzerland that provides an online magazine in easy language. The articles are mainly written in German but an increasing number of articles are translated into French. However, we only used the German articles for this corpus. *Infoeasy* addresses a variety of topics including international news, news about Switzerland, current topics, society, culture, health, sports, economy, and science.

3.3. Simplification Guidelines

In this work, the focus was on creating a corpus of simplified news that is as diverse and comprehensive as possible. The aim was to include texts from sources in several languages and on a broad variety of topics. As a result, also the level of simplification and the target group vary among the different news providers.

Informazione Facile and *The Times in Plain English* assess the complexity of their texts with standardized, length-based readability indices. For assessment, *Informazione Facile* uses the service of *corrige.it* that analyzes texts according to the GULPEASE Index (Lucisano and Piemontese, 1988). The GULPEASE Index is tailored to the Italian language and includes a scale that associates the index with a level of education. The platform states that their texts are suitable for people with reading skills at the level of basic school education. Additionally, *Informazione Facile* uses a basic vocabulary reference (Chiari and Mauro, 2014) to decide which words need further explanation. *The Times in Plain English*, on the other hand, uses the Flesch-Kincaid Grade Level, which assigns a school grade of the U.S. education system that is needed in order to be able to read the text at hand (Kincaid et al., 1975). However, the platform does not state the specific grade level used.

The magazines *Selkosanomat* and *Lätta Bladet* follow the guidelines for plain language that are listed on the website of Selkokeskus.[4]. They include instructions for the vocabulary, such as preferring well-known vocabulary and explaining difficult words, and for the language structure, such as writing short sentences and using active voice. Additionally, specific guidelines for different text types, including media texts, exist. These guidelines are developed and maintained by Selkokeskus.

No information is given by *Infoeasy* and *Journal Essentiel* as to which guidelines they use. Future work could empirically analyze the subcorpora and compare their usefulness for different target audiences.

3.4. Data Collection

To obtain all published articles, we developed web scrapers to scrape the archive pages of the providers. For this, all URLs of the articles were collected. By performing requests to the collected URLs, we obtained the HTML files of the article pages, which we then parsed. In the case of *Informazione Facile*, we received a database export in RSS format of the editors, therefore, no web scraping was needed.

To collect the newly published data, we use an RSS-based web scraping approach. Requests to the RSS feeds of the providers are made daily. The RSS files are parsed to extract the textual data and the metadata. In cases where not all information is contained in the

[4] https://selkokeskus.fi/selkokieli/

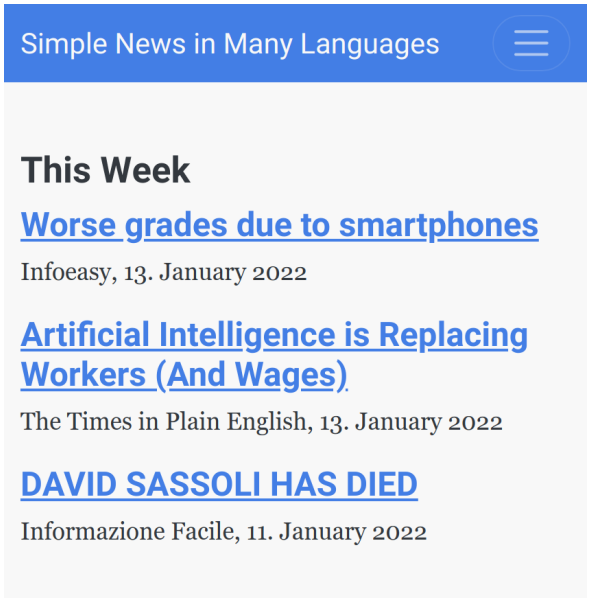

Figure 2: Screenshot of the web reader interface.

RSS file, the web pages of the articles are scraped to obtain the missing data. For each provider, we developed a specialized parser to parse the RSS and HTML files.

3.5. Dataset Format

The corpus is structured in XML. Each article is represented in an article element and can be identified with a unique ID. The textual data consist of the title, a short description and the complete text body of the news article. Additionally, metadata about the article are provided: the category or categories that the article was published in, keywords, the language, the URL, the publication date, the author, and the provider. Providers are further characterized with their name, a link leading to the website, the license, and a link to the license if one exists. As for now, the XML structure does not conform to the TEI format. However, we consider changing the format in a future version of the corpus.

Besides the complete corpus, we provide several subcorpora. To enable work on only one of the languages, a separate XML file is available for each provider. Also, files containing all articles published in a specific month are provided. For each month, additionally, a sub-corpus for each provider is compiled.

4. Web Reader Interface

In order to make the collected news articles not only available to researchers but also to the target groups of simplified language, we created a web-based reader interface in the style of a news aggregator. The user interface is available in German and English, while the news articles are provided both in their original language and (as machine translations) in the language of the user interface.

The articles are listed with their title, the original news provider, the publication date, and a short description of the content. They are ordered by publication date and are summarized under the week they were published in. The detail view of an article shows its complete text content.

5. Conclusion

The SNIML corpus compiles more than 13k simplified news articles in six languages. The articles are shared under an open license that permits academic use, and are planned to be continually updated.

The SNIML corpus is capable of serving as a useful complement to other resources for simplified language. For example, Simple English Wikipedia has grown very large but is restricted to a single language, text style, and simplification level. Complementing Simple English Wikipedia, the SNIML corpus could thus improve the diversity and multilinguality of language models for simplified language, as well as identification systems for simplified language.

Furthermore, the multilingual composition of SNIML opens up possibilities for the evaluation of cross-lingual transfer. For example, a system for the identification of simplified language could be trained on the English portion and evaluated on the five other languages in the corpus.

Moreover, the temporal stratification of the corpus makes it possible to evaluate a model for simplified language on concepts and topics unseen during training. It has been shown on the example of standard English that temporal generalization is a challenge for language models (Lazaridou et al., 2021), and we believe that it could even more so be a challenge for simplified language.

Future work may consist of aligning the articles to related articles in standard language. Such an extended version of SNIML could offer new opportunities for parallel corpus mining, or even for experiments towards an unsupervised text simplification system.

6. Acknowledgements

We thank the providers of the news articles in the corpus, namely *The Times in Plain English*, *Selkosanomat* and *Lätta Bladet*, *Journal Essentiel*, *Informazione Facile*, and *Infoeasy*, for agreeing to make their work available under an open license.

7. Bibliographical References

Aluísio, S. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles,

California, June. Association for Computational Linguistics.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., and Ebling, S. (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, May. European Language Resources Association.

Bott, S. and Saggion, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120, March.

Bredel, U. and Maaß, C. (2016). *Leichte Sprache. Theoretische Grundlagen. Orientierung für die Praxis*. Dudenverlag, Berlin.

Chiari, I. and Mauro, T. (2014). The new basic vocabulary of Italian as a linguistic resource. In *First Italian Conference on Computational Linguistics CLiC-it 2014*, pages 113–116, 10.

Gala, N., Tack, A., Javourey-Drevet, L., François, T., and Ziegler, J. C. (2020). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France, May. European Language Resources Association.

Grabar, N. and Cardon, R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands, November. Association for Computational Linguistics.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado, May–June. Association for Computational Linguistics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34.

Lucisano, P. and Piemontese, M. E. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31):110–124.

Maaß, C. (2020). *Easy Language – Plain Language – Easy Language Plus. Balancing Comprehensibility and Acceptability*, volume 3 of *Easy – Plain – Accessible*. Frank & Timme.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (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, July. Association for Computational Linguistics.

Rios, A., Spring, N., Kew, T., Kostrzewa, M., Säuberli, A., Müller, M., and Ebling, S. (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, November. Association for Computational Linguistics.

Saggion, H. (2017). *Automatic Text Simplification*. Morgan & Claypool Publishers.

Säuberli, A., Ebling, S., and Volk, M. (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, May. European Language Resources Association.

Wubben, S., van den Bosch, A., and Krahmer, E. (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, July. Association for Computational Linguistics.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 05.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.

8. Language Resource References

Centre Inria de Paris, Équipe ALMAnaCH. (2019). *OSCAR*.

Common Crawl. (2016). *CC-News*.

Appendix

Provider	Example	English machine translation
The Times in Plain English	With Omicron, you may get a scratchy throat, nasal congestion, a dry cough, and muscle pain in the lower back. These are the same symptoms as the Delta variant, and they are also the symptoms of the first coronavirus outbreak. An expert said, "It is still too early to say there is any difference in the Omicron symptoms."	-
Selkosanomat	Rokotuksia halutaan vauhdittaa, koska koronaviruksen uusi muunnos omikron leviää yhä nopeammin. Britanniassa, Tanskassa ja Norjassa omikron on levinnyt laajemmin kuin Suomessa. Näissä maissa suunnitellaan uusia tiukkoja rajoituksia.	-
Journal Essentiel	Je comprends les personnes qui disent: "Nous ne savons pas ce que ces vaccins pourraient nous causer à l'avenir." J'ai envie de leur répondre: "Il y a sans doute certains effets à long terme qui sont inconnus mais aujourd'hui, le vaccin est notre meilleur moyen pour sortir de cette pandémie."	*I understand people who say, "We don't know what these vaccines might do to us in the future." I want to answer them: "There are probably some long-term effects that are unknown but today, the vaccine is our best way out of this pandemic."*
Informazione Facile	• I bambini fino agli 11 anni riceveranno un terzo della dose prevista sopra i 12 anni. • La sperimentazione del vaccino è stata fatta su un piccolo numero di bambini: 2.300.	• *Children up to the age of 11 will receive one third of the expected dose over the age of 12.* • *The vaccine trial was done on a small number of children: 2,300,*
Lätta Bladet	Regeringen vill få fart på vaccineringen. I Finland finns det fortfarande ungefär 800 000 vuxna som inte har fått coronavaccin. Regeringen har också bestämt att man nu börjar vaccinera barn över 5 år mot corona.	-
Infoeasy	Genesen ist ein anderes Wort für: wieder gesund. Wir brauchen darum jetzt an vielen Orten ein Covid-Zertifikat. Nur mit dem Zertifikat dürfen wir hinein. Und dieses Zertifikat bekommen nur Personen, • die geimpft sind. • die genesen sind. • die einen Corona-Test gemacht haben. Und der Test muss negativ sein.[5]	*Recovery is another word for: healthy again.* *That's why we now need a Covid certificate in many places.* *We are only allowed in with the certificate.* *And this certificate is only given to people* • *those who are vaccinated.* • *that have recovered.* • *who have taken a corona test. And the test must be negative.*

Table 3: Text examples for each news provider. Machine translations are provided if the license permits derivative work.