Department of Digital Humanities
University of Helsinki
Dissertationes Universitatis Helsingiensis 232/2025

# Resources and Tools for Automatic Text Simplification: Cases of Russian and Finnish

Anna Dmitrieva

ACADEMIC DISSERTATION
To be presented, with the permission of the Faculty of Arts of the University of Helsinki, for public discussion on the 17th of June 2025 at 15 o'clock. The defense is open for the audience through remote access.

Helsinki 2025

**Pre-examiners**

Professor Horacio Saggion, Universitat Pompeu Fabra

Associate professor Liana Ermakova, University of Brest

**Custos**

Professor Jörg Tiedemann, University of Helsinki

**Supervisors**

Professor Jörg Tiedemann, University of Helsinki

Docent Ulla Vanhatalo, University of Helsinki

Professor Roman Yangarber, University of Helsinki

**Opponent**

Associate professor Liana Ermakova, University of Brest

Publisher: Helsingin yliopisto
Series: Dissertationes Universitatis Helsingiensis 232/2025
ISBN 978-952-84-1063-8 (Paper back)
ISBN 978-952-84-1062-1 (PDF)
ISSN 2954-2898 (Print)
ISSN 2954-2952 (PDF)
PunaMusta, Joensuu 2025

# Abstract

Automatic text simplification is a natural language processing task for making a text more readable and accessible to a broader audience while preserving most of its informational content. Today, text simplification is viewed as a monolingual machine translation task and is mostly done by training neural network language models. There exist many methods for building simplification models, but not a lot of works deal with less-represented languages such as Russian or Finnish. The goal of this doctoral project was to fill this gap by creating new data sources for these languages and training simplification models.

Simplification models typically require a lot of parallel data to train on. A substantial part of this work deals with creating parallel datasets from raw text data. During my studies, I have assembled parallel datasets for Russian and Finnish. In these datasets, each piece of text in “standard” language has a corresponding simplified version. All datasets have sentence-aligned versions, since this is the default method of alignment for automatic text simplification. The Russian datasets include mostly literary texts such as classical Russian literature and fairy tales, as well as an encyclopedic subcorpus. The Finnish datasets are based on news articles. In all cases, simplification was originally performed by professionals. These parallel datasets are available online and can be used for training simplification models or for studying text simplification strategies that experts use, as well as the linguistic features of simplified texts.

Since automatic text simplification is considered similar to a machine translation task, it is solved with similar methods. At present, it is mostly approached by fine-tuning large language models. Most of the architectures used for simplification are encoder-decoder or decoder-only neural networks with attention. Over the course of this thesis project, I trained sequence-to-sequence models from scratch, as well as fine-tuning larger models such as the multilingual BART, T5, and the Finnish GPT. I also experimented with multiple modeling strategies, including multi-task learning, controllable simplification, and instruction fine-tuning. These techniques can, for example, aid in transferring knowledge from data-rich to low-resource tasks or tailoring the model's output to fit the needs of a specific audience.

The models' outputs have been evaluated automatically with language-agnostic metrics for simplification quality evaluation, such as the SARI score. In some cases, the scores could not yet be compared to any other works: for example, in the case of Finnish simplification, to the best of our knowledge, there were no other works to compare our results to at the time of our experiments. In other cases, the models showed performance comparable to the best results previously achieved on similar data and tasks. In addition to automatic evaluation, some outputs have been evaluated empirically, and their linguistic features, such as the level of sentence compression, the proportion of additions and deletions, etc. were examined.

# Acknowledgements

I became interested in automatic text simplification in 2013, when I applied to participate in a research group on this topic. When it was time to choose a topic for my PhD, simplification naturally came up: my interest in this field and the concept of simplicity itself never went away. I am grateful to everyone who made my work on this project possible, even if I may not be able to name each of them in this limited space.

First and foremost, I would like to thank my supervisors: Prof. Jörg Tiedemann, Docent Ulla Vanhatalo, and Prof. Roman Yangarber. The list of ways they have supported me over the years could not be simplified without a significant loss of information. My thanks also go to my thesis committee members, Docent Lidia Pivovarova and Prof. Camilla Lindholm, for their valuable advice and for always providing a fresh perspective on my work. I thank all my co-authors, especially Maria Lebedeva and Antonina Laposhina, for the opportunity to work with them and learn from them. I am also deeply grateful to my pre-examiners, Prof. Liana Ermakova and Prof. Horacio Saggion, for reading this dissertation and for their suggestions on improving the text.

This thesis is largely based on three individual research projects on automatic text simplification for Russian and Finnish texts. I would like to thank the Finnish National Agency for Education (EDUFI), the Juhlarahasto (Cultura Foundation), and the Finnish Cultural Foundation (SKR) for supporting my work on this thesis.

Finally, I am grateful for all the love and support that I received and continue to receive from my family and friends. I want to thank Aleksandra Konovalova and Anton Lakstygal for making my stays in Finland a thousand times better than they would have been without them. As my co-author, Aleksandra is also the person who made the Finnish-Easy Finnish parallel corpus happen and helped with many other Finnish-related things in my work, studies, and everyday life. Above all, I am grateful to my parents, Svetlana Dmitrieva and Andrei Dmitriev, my grandmother, Svetlana Dmitrieva, and my husband, Sergei Korobenkov, for believing in me. Without them, even the simplest things would not have made sense.

Назавтра лекцию свою он начал с достоинства слога. [...] Простой слог был способ писать так, как говорят. Некоторые его ложно называли низким только потому, что он не был высок. Но выражения, слова, мысли в сем слоге вовсе не низки, они обыкновенные, но благородные.

Ю.Н. Тынянов, “Пушкин”

The next day, he began his lecture with the topic of the dignity of style. [...] Simple style is a way of writing as one speaks. Some falsely consider it “low” only because it is not elevated. But the expressions, the words, and the thoughts in this style are not low at all; they are ordinary, yet noble.

Yury Tynyanov, “Pushkin”

# Index

## Errata

On page 59, Section 4.2, the equation should be:

$$SARI = \frac{1}{3}F_{add} + \frac{1}{3}F_{keep} + \frac{1}{3}P_{del}$$

# List of original publications

This thesis is based on the following publications:

I. Dmitrieva, A., & Tiedemann, J. (2021). A Multi-task Learning Approach to Text Simplification. In: van der Aalst, W.M.P., et al. Recent Trends in Analysis of Images, Social Networks and Texts. AIST, 2020. Communications in Computer and Information Science, vol 1357. Springer, Cham. The final authenticated publication is available online at DOI: https://doi.org/10.1007/978-3-030-71214-3_7

   This work is about combining summarization and simplification methods to achieve better simplification results. The research was done on English datasets, since at the time of working on this paper, there were no publicly available Russian simplification datasets. We explored methods that would allow for the transfer of knowledge from other tasks, such as summarization, for successful automatic text simplification, as well as methods for working in low-resource settings.

   Author contributions: AD & JT: conceptualizing the study. AD: data collection, data preprocessing, modeling, and writing the paper. JT: consultations on methods, coordinating the experiments, and editing.

II. Dmitrieva, A., & Tiedemann, J. (2021). Creating an Aligned Russian Text Simplification Dataset from Language Learner Data. n Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 73–79, Kyiv, Ukraine. Association for Computational Linguistics. URL: https://aclanthology.org/2021.bsnlp-1.8

   This paper describes the creation of a Russian-Simplified Russian parallel dataset. At the time of publication, the dataset comprised literary texts adapted for learners of Russian as a foreign language and their original versions. The texts were aligned on the paragraph level. This paper also includes an experiment on automatic Russian simplification with an LSTM model.

Author contributions: AD: obtaining and processing the data, modeling, and writing the paper. JT: supervising the dataset creation and experiments, editing.

III. Dmitrieva, A., Laposhina, A., & Lebedeva, M. (2021). A Comparative Study of Educational Texts for Native, Foreign, and Bilingual Young Speakers of Russian: Are Simplified Texts Equally Simple? Frontiers in Psychology, 12, 703690. doi:10.3389/fpsyg.2021.703690

The study investigates the differences between simplification strategies used in texts for different groups of young Russian language learners. We used various classification models to compare text complexity and simplification strategies between three different domains (texts for native, foreign, and bilingual young speakers). We also described some linguistic properties of the simplified texts.

Author contributions (quoting the paper): AD: literature review, conducting preprocessing, classification model building and evaluation, data analysis, and writing the paper. AL: data collection and annotation, data analysis, interpretation of results, and writing the paper. ML: conception and design of the study, formulation of research goals and aims, literature review, interpretation of results, writing the paper, overall management, and coordination of the study. All authors contributed to the article and approved the submitted version.

IV. Dmitrieva, A., Laposhina, A., & Lebedeva, M. Y. (2022). Creating a list of word alignments from parallel Russian simplification data. Frontiers in Artificial Intelligence, 5, 984759. doi:10.3389/frai.2022.984759

This paper follows the creation of a list of word alignments based on the literature sub-corpus of the RuAdapt dataset. The list was created to study the strategies of lexical simplification used by experts and to aid in developing educational materials for teaching Russian as a foreign language. The word pairs were collected automatically and assessed by multiple experts.

Author contributions (quoting the paper): AD: conception and design of the study, formulation of research goals and aims, data collection, literature review, aligners building and evaluation, and writing the article. AL: literature review, data analysis, interpretation of results, and writing the article. ML: conception and design of the study, coordination of the expert

annotation process, interpretation of results, and writing the article. All authors contributed to the article and approved the submitted version.

V. Dmitrieva, A. (2023). Automatic text simplification of Russian texts using control tokens. In Proceedings of the 9th Workshop on Slavic Natural Language Processing, 2023 (SlavicNLP, 2023), pages 70–77, Dubrovnik, Croatia. Association for Computational Linguistics. URL: https://aclanthology.org/2023.bsnlp-1.9

This paper deals with the controllable simplification of Russian texts. The controlled dimensions are length, paraphrasing degree, syntactic complexity, and the CEFR (Common European Framework of Reference) grade level of the text. These attributes are manipulated using special control tokens added to the beginning of the input sentence.

VI. Dmitrieva, A. (2023). The role of language technology in accessible communication research. In: Deilen, S., Hansen-Schirra, S., Garrido, S., Maaß, C., Tardel, A. (eds) Emerging Fields in Easy Language and Accessible Communication Research. Easy – Plain – Accessible, vol 14. Frank & Timme, Berlin. DOI: https://doi.org/10.57088/978-3-7329-9026-9_12

This is an overview paper of ways in which accessible communication studies interact with language technology research. It tackles various dimensions of communication and language, such as automatic text simplification, speech-to-text and text-to-speech translation, and text-to-pictogram translation. The goal is to show how these technologies aid in facilitating accessible communication.

VII. Dmitrieva, A., & Konovalova, A. (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. URL: https://aclanthology.org/2023.crowdmt-1.3/

In this article, we describe the development of the first Finnish-Easy Finnish parallel dataset. The texts in the dataset were taken from the general Yle archives for 2019-2020 and Yle Uutiset Selkosuomeksi archives for the same years, available on Kielipankki (The Language Bank of Finland). The articles were first aligned automatically, and then the pairs with good similarity scores were evaluated by a human expert. The resulting dataset is now available on Kielipankki.

Author contributions: AD&AK: conceptualizing the study, reviewing relevant literature, and writing the paper. AD: data collection and preprocessing, automatic document alignment. AK: manual evaluation of the document pairs, analysis of the linguistic features of the simplified articles.

VIII. Dmitrieva, A., & Tiedemann, J. (2024). Towards Automatic Finnish Text Simplification. In Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING, 2024, pages 39–50, Torino, Italia. ELRA and ICCL. URL: https://aclanthology.org/2024.determit-1.4/

This paper describes the augmentation of the Finnish-Easy Finnish dataset and the making of baseline Finnish simplification models. Articles from the years 2014 to 2018 have been aligned on the document level, and sentence-aligned versions of all document pairs from 2014 to 2020 were added. For automatic Finnish text simplification, we experimented with two models: mBART and Finnish GPT XL.

Author contributions: AD: dataset augmentation (including data collection, choosing the optimal aligners, and creating new alignment solutions where necessary), modeling, and writing the paper. JT: supervising the experiments, consultations, and editing.

The publications are referred to in the text by their Roman numerals.

# 1 Introduction

## 1.1 Motivation

This thesis focuses on tools and resources for automatic text simplification (ATS), particularly simplification for languages with few or no specially designed resources for simplification. In this work, I will describe all the steps in creating a data-driven simplification system, from assembling a parallel dataset to creating a simplification model and assessing its quality. I focus on languages less represented in this field, namely Russian and Finnish.

Text simplification (TS), defined narrowly, is the process of reducing the linguistic complexity of a text, while still retaining its original information content and meaning (Siddharthan, 2014). Automatic text simplification is a task of natural language processing (NLP). The "simple language" in the context of this work can be thought of as a linguistic variety that falls on the spectrum of easy-to-understand varieties such as Easy and Plain Language. These and other comprehensibility-enhanced language varieties are all part of accessible communication (Maaß, 2020).

Nowadays, accessible communication is given a high priority in many countries (Maaß, 2020). For example, some European governmental organizations provide accessible versions of their text materials and websites - such as Finnish websites in Easy Finnish. The target audience of these comprehensibility-enhanced materials can be vast and include anyone who does not possess the literary skills of an educated adult native speaker or simply does not have expertise in a particular field (for example, law or medicine). A large number of consumers of these media is also made up of people from marginalized groups, such as migrants or people living with a disability.

The field of automatic text simplification dates back to the middle of the 1990s (see, for instance, Chandrasekar and Bangalore, 1997, and Chandrasekar et al., 1996). Early work on this topic aimed mainly at grammar and style simplification in English. In the beginning, text simplification systems were mostly rule-based. Today, text simplification is most often viewed as a monolingual translation problem (Siddharthan, 2014). Modern text simplification tools are created using neural machine translation (NMT) and text generation techniques. Using deep neural networks has proven to allow for the creation of simple (as in correlating

with the human judgment of simplicity) and fluent modifications that preserve the meaning of the original sentence (Zhang and Lapata, 2017).

Modern deep neural networks that solve monolingual sequence-to-sequence translation and text generation tasks are trained and/or fine-tuned on large amounts of data and learn a desired task or set of tasks in context from the given training data. For simplification, it means that, in most cases, a model learns this task by training on parallel datasets where each "regular" text has a corresponding simplified version. The more good-quality data is available, the better the quality of the simplification will be. More extensive data sources also open up possibilities for more tasks: for example, simplifying a text to a certain complexity level, doing explanatory simplification where complex concepts are augmented with short definitions, etc. Even the introduction of large conversational AI models such as ChatGPT has not eliminated the need for training data but rather increased it.

The vast majority of new simplification techniques and datasets are created for and/or tested on English data in the first place (see, for example, Alva-Manchego et al., 2020, for a TS benchmark dataset, or Maddela et al., 2023, for a state-of-the-art simplification evaluation metric). In recent years, the situation has started to change, and a lot of new non-English parallel datasets have appeared (Ryan et al., 2023). However, many languages still do not have their own simplification datasets. This has led to the field of ATS advancing in two different branches. On the one hand, the number of solutions for the efficient creation of monolingual parallel corpora (i.e., corpora that consist of pairs or longer sets of texts with equivalent meaning), such as automatic sentence aligners, is increasing (see, for example, Thompson and Koehn, 2019, or Štajner et al., 2018). On the other hand, researchers are looking for means of enhancing the simplification models' performance. This includes, among other research tasks, finding ways to transfer knowledge between languages and tasks - for example, multi-task learning, where a model would learn several tasks simultaneously (summarization and simplification, for instance) and use the knowledge of the more-represented tasks to aid in solving the less-represented ones, like using machine translation models and frameworks for text simplification (see Cooper and Shardlow, 2020 as an example).

As previously mentioned, there exist different levels of enhanced comprehensibility, and the target audience of accessible communication can be very broad. It does not seem possible to create a separate model for each group of users and for each complexity level. Instead, a model can be trained in a way that allows for controlling certain linguistic properties of the output. These can be simple things such as length or more complicated properties such as modality or complexity level. Controlled simplification is another rapidly developing sub-field of TS.

Despite the growing body of research on TS, it is still a challenging task in natural language processing. For example, automatic evaluation of simplification systems output remains an open research question, since current best practices do not take

grammaticality or cohesion into account, and the emerging data-driven approaches are not language-agnostic. In my own research, I focused on building datasets and simplification models for less-represented languages in TS. Over the course of my project, I also studied various strategies that can help enhance the simplification performance, such as multi-task learning, data augmentation, and controlled simplification. My goals were to find ways of performing efficient simplification in limited data settings, provide open-access datasets for researchers on text simplification, accessible communication, and other fields, and study the linguistic properties of simplified language.

## 1.2 Research Objectives and Thesis Structure

Since this thesis project is focused on non-English text simplification, the two main research objectives of this study were to:

1. Create new simplification-specific data sources for languages that lacked such datasets, namely Russian and Finnish. Make these resources accessible at least for academic use.
2. Develop methods to perform automatic text simplification for these languages. Train simplification models of various kinds and assess their output quality.

The initial hypothesis of this study was that, much as there are ways to perform multilingual machine translation in low-resource settings, it is also possible to develop simplification systems for languages with limited resources, specifically Russian and Finnish, by creating new parallel datasets and exploring training strategies adapted to low-resource settings.

The main research questions explored in this thesis are:

1. What does it mean to simplify a text? What linguistic properties characterize a simple text, and how can simplicity be evaluated? [RQ1]
2. What are the most effective strategies for creating parallel datasets for text simplification from various data sources and domains? [RQ2]
3. Can simplification models be trained to achieve optimal performance when dealing with limited data in morphologically rich languages? [RQ3]

The next chapter in this thesis, Chapter 2, deals with the terminology issues of this study and examines the linguistic properties of simple and simplified texts. It incorporates Papers III and VI, which explore the writing strategies in simple texts and the links between text simplification as a task of natural language processing and accessible communication studies, respectively. Chapter 2 also includes a comparative study of several Easy and Plain Language guidelines, which has not been incorporated into the published papers of this thesis. Generally, this chapter defines simplicity and simplification within the context of this work.

Chapter 3 describes the process of creating various data sources for automatic text simplification and studying the linguistic properties of simplified texts. Perhaps the most important practical contributions of this thesis are the two parallel simplification datasets for Russian and Finnish created and released over the course of working on this project. Chapter 3 incorporates Papers II and IV, which describe the development of the parallel Russian-Simple Russian dataset (RuAdapt) and a supplementary parallel word list based on the RuAdapt texts. This chapter also refers to Papers VII and VIII, which outline the creation of the parallel Finnish-Easy Finnish dataset.

Chapter 4 describes the experiments on training text simplification models on various data and evaluating the models' output. This chapter, as well as Chapter 3, includes content from Papers II and VIII, but this time to describe training and fine-tuning simplification models on the newly developed datasets discussed in Chapter 3. It also includes specific training techniques such as transfer learning with the intention of using it for low-resource simplification (Paper I) and experiments with controlled simplification where the output can be tailored for a specific audience (Paper V).

Finally, Chapter 5 concludes this thesis, revisiting the research questions, summarizing the contributions of the present study, and giving a perspective on possible future work.

# 2 Background

## 2.1 Easy and Plain Language, Simple and Simplified Language

In this chapter, I will discuss the terminology necessary for understanding the content of this thesis. This chapter addresses the first research question (RQ1) by describing what it means to simplify a text and what the expected outcome of the simplification process can be. Moreover, in this chapter I discuss the linguistic properties of simple language and the writing strategies used by authors of simple texts, especially in relation to Easy/Plain Language guidelines. Some of the writing strategies that exist in simple and simplified texts were studied in Paper III using Russian textbooks for young speakers in various language environments. In addition, I will refer to Paper VII to discuss how the linguistic properties of Easy Finnish news articles correspond to the Selkomittari – a set of criteria for assessing Easy Finnish texts. This chapter also touches upon the interaction between language technology and accessible communication research, as per Paper VI.

This thesis exists between two fields that rarely intersect: accessible communication (AC) studies and monolingual machine translation. I will first talk about accessible communication since it provides the framework of definitions for all things related to comprehensibility enhancement in language. Defining terminology in the context of this work presents a complex challenge for many reasons. First of all, giving precise definitions to abstract terms such as “accessibility", “understandability", and “simplicity/complexity” has proven to be a rather nuanced undertaking. Secondly, the terminology for different degrees of simplicity and accessibility, such as Easy/Plain Language, varies across different languages. It should be noted that I am limiting myself strictly to describing the terminology in the scientific community and not the connotations that these terms might have in everyday use; i.e. I will not address the acceptability of these terms among the general public.

One of the definitions of the word “accessible” given by the Merriam-Webster dictionary[1] reads as “easily used or accessed by people with disabilities: adapted for use by people with disabilities.” However, accessible communication is not only for people with disabilities. According to Perego (2020), accessibility is a universally

[1] https://www.merriam-webster.com/dictionary/accessibility

inclusive concept that is based on the idea of availability and is not necessarily linked to persons with disabilities but applied to all while hingeing on the general ability to use products or services. Hence, the expression "accessible communication" refers to any form of simple or simplified communication that prevents communicative exclusion (ibid.).

It has been established in the AC research community that there is a spectrum of enhanced comprehensibility or reduced complexity in communication. In general, the term "Easy Language" denotes the most understandable variety of language, and "Plain Language" is something between Easy Language and standard language (see, for example, Maaß, 2020). In German, Leichte Sprache ("Easy Language") is conceived as a firmly rule-based variety with clear outlines, whereas Einfache Sprache ("Plain/Simple Language") is seen as a continuum ranging from somewhat enriched forms of Easy Language to forms somewhat below average standard German or languages for special purposes (like legal or medical communication) (Maaß, 2020). In Russian, there is a similar distinction between the terms "простой язык" (Plain Language) and "ясный язык" (Easy Language). In Finnish, Easy Language is "selkokieli," Easy Finnish is "selkosuomi" and Plain Language is "selkeä kieli" or "selkeä yleiskieli" (Leskelä, 2021); however, "selkokieli" and "selkosuomi" appear to be used in almost all cases when describing comprehensibility-enhanced Finnish. Sometimes, the terms Easy-to-read or Easy-to-understand are also used as equivalent terms for Easy and Plain Languages (for instance, Maaß (2020) states that "Easy-to-read" is the most frequently used English equivalent of Easy Language). However, "Easy-to-read" refers to a quality of written texts that makes it easy to extract the content (Maaß, 2020), "Easy-to-understand" can go beyond written forms of communication.

Still, when we see a text "in the wild", be it an adapted text for children or second language learners, a simplified newspaper article, or just a short and readable text, it is not easy to say definitively whether or not this text was written in Easy, Plain, or some other variety of comprehensibility-enhanced Language. It can be established that the text is written in Easy Language if it adheres to certain criteria (since Easy Language tends to be the most regulated form of accessible communication); however, the criteria are different for each language. For example, to the best of our knowledge, it has never been established that any text adapted for second language learning on the elementary level (CEFR[2] A1) can or may be used as an example of Easy or at least Plain Language, although, intuitively, it may. In this work, I will also use terms such as "adapted" or "simplified". Both of them will

[2] Common European Framework of Reference for Languages is a detailed model for describing and scaling language use and the different kinds of knowledge and skills required (Council of Europe et al., 2001). It is a widely used framework for assessing and comparing language skills across different languages and educational systems.

refer to texts that have undergone specific changes, made either by a human expert or a language model, that were aimed at making the texts easier to understand.

It should be noted that the terms “simplified text” or “simple text” used in this work do not mean the same thing as, for example, Simple English in Simple English Wikipedia. SimpleWiki articles are created using a set of criteria for making them easier to understand[3], which includes suggestions on writing in Basic English[4], which makes them, in a way, examples of a regulated language variety. The term “simple” in this work is also not used as a synonym for “Plain [Language]". Sometimes the terms “plain” and “simple” are used interchangeably, particularly when referring to languages where the word for “Plain” as opposed to “Easy [Language]” also means “simple” (like “einfach” in German or “простой” in Russian). I use the terms “simple” and “simplified” in a broader sense, often in opposition to non-simplified, original texts. For example, if a neural network translates a CEFR C2 text into a CEFR C1 text, we will refer to the input text as original and to the output text as simple or simplified, even though it may still not be very easy to understand. For the purposes of this research, any text, be it adapted for a certain group of language users (children, second language learners, people with learning disabilities, etc.), translated into Plain or Easy Language, or in any other way modified to be easier to understand, can be called simplified. At the same time, any text that is written or created with the intention of making it easy to understand can be called simple. However, it should be noted that a simple text is always created following certain rules, of which the creator is aware either consciously (e.g., a second language teacher creates an exercise for her students based on what lexis and grammar they already know), (semi)-intuitively (e.g., a writer composes an article in her native language using common vocabulary to appeal to a broader audience), or from the data it has been based or trained on if the creator is a language model. To sum up, all simplified texts should be simple (or at least simpl**er** than the texts they have been derived from), but not all simple texts have an “original” version from which they have been derived by simplification.

## 2.2 Text Simplification and Accessibility

In Paper VI, I describe some ways in which accessible communication research crosses paths with language technology.

As mentioned before, automatic text simplification has been researched since the 1990s. In 1996, Chandrasekar et al. published their work on syntactic sentence simplification (Chandrasekar et al., 1996). They aimed to simplify the input for

[3] https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

[4] https://simple.wikipedia.org/wiki/BASIC_English

natural language applications that existed at the time. Earlier TS systems were rule-based and often used formal grammars to represent sentences and rewriting operations (Chandrasekar and Bangalore, 1997; Dras, 1999).

Earlier in the development of the field, lexical (Paetzold and Specia, 2017) and syntactic (see Brouwers et al., 2014; Scarton et al., 2017) simplification were often treated as separate tasks. However, with the rise of data-driven approaches in NLP, hybrid techniques that involved lexical and syntactic simplification, as well as discourse-level simplification, such as content reduction, became more common. For example, the best results for ATS on the Amazon Mechanical Turk benchmark (Xu et al., 2016) have been achieved by models such as GPT-3 (Brown et al., 2020) and mBART (Liu et al., 2020)[5]. These large language model architectures learn in context, which means that they learn desired behaviors, such as simplification, by training on a large number of example pairs of regular and simplified sentences. That means that, theoretically, any simplification strategy that occurs in the dataset can be learned by the model.

TS applications can have a very broad target audience, and many of its target groups can be found among people who could benefit from using Easy and Plain Language materials. These target groups include deaf and hard-of-hearing people (Alonzo et al., 2020), people with cognitive impairments (Espinosa-Zaragoza et al., 2023), readers with low literacy (Aluísio and Gasperin, 2010), children (De Belder and Moens, 2010), and foreign language learners (Degraeuwe and Saggion, 2022). In other words, these are the people whose literacy skills are below those of an educated adult native speaker not living with a disability. However, an "average" person can also benefit from simplified texts when she is dealing with expert language. For example, TS applications also exist for medical (Devaraj et al., 2021) and legal (Garimella et al., 2022; Pereira et al., 2024) texts.

It should be noted that text is not the only mode of communication where language technology can aid understanding. For example, speech-to-text and text-to-speech technologies (see Matamala, 2016; Bouillon et al., 2021) can also bridge communicative gaps, especially for visually impaired users. Right now, speech-to-sign and sign-to-text technologies are also experiencing a boost (see, for example, Shterionov et al., 2023). Moreover, there are text-to-pictogram applications that aid communication in cases where text or speech either cannot be used or needs significant augmentation (see Vandeghinste et al., 2017; Vaschalde et al., 2018).

Overall, it is evident that various NLP technologies can help facilitate accessible communication in different domains. However, it is also clear that simplification has not yet achieved the popularity of multilingual machine translation among average users. There is no simplification app as widely used as, for example, Google

[5] https://paperswithcode.com/sota/text-simplification-on-turkcorpus. At the time of writing of this thesis, the best model is GPT-3-175B (Vadlamannati and Şahin, 2023), and the second best is MUSS (Martin et al., 2022).

Translate is for translation, and no popular browser extension that would simplify a website's content in one click. The closest alternative to a go-to simplification app for the general public nowadays has to be large multi-purpose language models such as ChatGPT or Bing Chat. ChatGPT shows good simplification results on the general domain (Li et al., 2023) and promising, although not perfect, results for some specific genres such as radiology reports (Jeblick et al., 2023; Lyu et al., 2023). However, it is unclear which languages besides English can be correctly simplified by ChatGPT. For example, studies have shown that ChatGPT performs worse on Portuguese domain-specific simplification than on German simplification for the same domain (Herget and Alegre, 2023), that domain being environment and climate change. It is also unclear whether it is able, for instance, to alternate consistently between different complexity levels or translate into Easy Language following language-specific instructions. Nevertheless, the introduction of large multi-task models has revolutionized the simplification field in that it has given the average user the opportunity to obtain good (or at least decent) quality simplifications quickly in a convenient interface. It is likely to have a great impact on both fields of accessible communication and text simplification, although it is yet unclear how exactly those fields will be influenced.

## 2.3 Linguistic Properties of Simple Language

As stated in Chapter 1, nowadays, accessible communication is treated as a high priority in many countries (Maaß, 2020). Many media get translated into Easy or Plain Language: examples from the Nordic countries include Easy Finnish news Yle Uutiset selkosuomeksi[6] and the Finnish Social Insurance Institution Kela's website in Easy Language[7], the Easy Swedish news 8 Sidor[8], and books in Easy Swedish adapted and published by the private publisher LL-förlaget[9] (the latter also publishes books that were not adapted from "regular" Swedish but originally written in Easy Swedish). Therefore, many languages now have guidelines for Easy or Plain Language translation. In order to better understand the linguistic properties of simplified texts, I will present an analysis of some of these guidelines and identify common features in them. The guidelines that I examine in this work are not state-mandated but rather have been created by professionals in the field to offer guidance on text adaptation.

Naturally, the guidelines are prescriptive rather than descriptive, as they state what an author should do or what a text should look like instead of what a text in Easy or Plain Language usually looks like. Most guidelines do not explain why a

[6] https://yle.fi/selkouutiset
[7] https://www.kela.fi/web/selkokieli/selkokieli
[8] https://8sidor.se/
[9] https://ll-forlaget.se/

certain rule is implemented, i.e., why a certain change will simplify the text. However, many of the rules are self-explanatory, such as the rules of shortness, familiarity, and commonness of words. Other rules seem to have been derived from experience. Given also the fact that simple languages are highly regulated language varieties, it is safe to assume that the Easy/Plain Language guidelines cover most linguistic features that are actually present in simple texts.

Some studies have focused on empirical evaluation of the recommendations typically provided in Easy/Plain Language guides. For example, (González-Sordé and Matamala, 2023), compiled a literature review of multiple guides and empirical studies. They found studies that evaluated (1) visual support, (2) linguistic simplification, (3) word frequency, (4) literacy mediation, (5) connectives, (6) coreferences, and (7) number of sentences, text length, and word length (ibid.). They also stated that some of this research was inconclusive, and the state of empirical research on Easy Language guideline recommendations is incipient.

For Finnish, there exists a set of criteria for assessing Easy Finnish texts called the Easy Finnish Indicator or Selkomittari[10]. At the time of writing, the second release is the most up-to-date version. The Indicator was created by the Finnish Centre for Easy Language. It should be noted that the authors of Selkomittari deem it not suitable to use as a guideline for writing in Easy Finnish. However, it does list the linguistic criteria that a text should meet in order to be considered written in Easy Finnish.

Selkomittari 2.0 has a total of 96 criteria for easiness, which are classifiedas follows: Text as a whole (27 criteria), Words (16 criteria), Language structures (24 criteria), and Layout and illustrations (29 criteria). Since this thesis is focused on text per se and not on media products in simple language, the latter group of criteria is not relevant to the current discussion.

In Russia, Easy and Plain Language does not yet have an official status, and there are no state-produced media in simplified language. Some manuals for social workers include recommendations on communicating with people with intellectual disabilities that resemble Easy Language guidelines (Nechaeva et al., 2020). A comprehensive guideline on Easy Russian with a focus on people with disabilities has been produced in Belarus (Khitriuk et al., 2018). It covers preferred vocabulary and sentence structures, as well as typographic rules, document layout, and infographics.

Although this thesis deals mainly with Russian and Finnish simplification, I decided to compare the guides available in these languages to the most often used and/or extensive guides for other European languages.

The first guide that I chose for comparison was “Information for all: European standards for making information easy to read and understand” (Inclusion Europe,

[10] https://selkokeskus.fi/in-english/easy-finnish/the-easy-finnish-indicator-2-0/

2010). It was developed by Inclusion Europe[11] and has been translated into multiple languages. I have used the English version; however, there is little language-specific information in the guide. Many Easy Language researchers and translators refer to these guidelines, with some even stating that they are or used to be the only available guide of this kind for their primary language.

The other European-wide multilingual guide that I used was "How to write clearly" by the European Commission (European Commission, 2015). This guide has been translated into more languages than Information for all. Unlike the previous guide, it is oriented more towards Plain Language rather than Easy Language suitable for people with disabilities.

In addition to the language-agnostic guides, I also picked three language-specific ones. I looked for more extensive guides and chose "Communiquer pour tous" for French (Ruel et al., 2018), "In duidelijk Nederlands" for Dutch (Taaltelefoon, 2012), and "Die Regeln für Leichte Sprache" for German (Netzwerk Leichte Sprache, 2022).

It should be noted that not all of these guidelines fall into the same place on the spectrum of Easy to Plain to "standard" language. For example, the European Commission's guide is aimed more towards the general audience and writing in plain language, whereas the Inclusion Europe guide, having been made with the needs of people with disabilities in mind, leans more towards Easy Language. The Russian, Finnish, and German guides are also made for Easy Language. The French guide is all-purpose, with some advice marked as being "for people with disabilities": these instructions are for a higher degree of simplification than the rest. The Dutch guide is geared towards the public domain and everyday communication, making it more of a Plain Language guide.

Usually, the guides are separated into sections for different text levels: words, sentences, text as a whole, etc. In my comparison, I have identified the following main sections: Words and phrases, Numbers, Formatting, Sentences, Structure, and Tone. The majority of the rules listed in the tables below, except for the most general ones (such as "Short words"), were taken directly from the guides; sometimes, the wording has been slightly changed or shortened. In rare cases, a rule could have been attributed to a different section than in the original guide: for instance, what was in the "Words" section could have been moved to "Tone".

In all tables below, "IE" stands for "Information for All" (Inclusion Europe, 2010), and "EC" stands for "European Commission" (European Commission, 2015).

[11] https://www.inclusion.eu/

### 2.3.1 Words and phrases

**Table 2.1.** Rules for words and phrases.

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Words and phrases** | | | | | | | |
| General | | | | | | | |
| Short words | | | + | | + | + | + |
| Easy-to-understand, well-known | + | + | + | | + | + | + |
| Less formal words | | | | + | | | |
| Less difficult words | + | | | + | + | | |
| Grammar | | | | | | | |
| *Morphemes* | | | | | | | |
| The text contains no words that have several different elements, such as derivative affixes, inflectional suffixes, and clitics | | | | | | + | |
| *Verbs* | | | | | | | |
| Where possible, use the present tense rather than the past tense | + | | + | | | + | |
| Conjugate the verb in the same tense throughout the document | | | + | | | | |
| Moods used are mostly the indicative (I listen, we speak), and the second person imperative (listen, speak) | | | | | | + | |
| The conditional is only used if it can't be replaced by an indicative without the clearly changing the meaning | | | + | | | + | |
| Use imperative when giving instructions | | | | + | | | |
| Prefer active verbs to passive ones | + | + | + | + | + | + | + |
| Use the positive form, not the negative | + | + | + | | + | + | + |
| Avoid using participial and adverbial phrases | | | | | + | | |
| Avoid the subjunctive | | | | | | | + |
| *Nouns* | | | | | | | |
| Be careful with nominalizations | | + | + | + | | + | + |
| Nouns should not have complex modifiers such as participle structures | | | | | | + | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Words and phrases** | | | | | | | |
| The text uses the basic forms of nouns if it is possible and natural in the context | | | | | | + | |
| The text mainly contains the easiest forms of noun inflections | | | | | | + | |
| Avoid the genitive | | | | | | | + |
| Always write the male form first | | | | | | | + |
| Anaphora avoidance and unambiguity | | | | | | | |
| Avoid ambiguity | | + | + | | + | + | |
| Use the same word to describe the same thing throughout your document | + | + | + | + | + | + | + |
| Be careful when using pronouns such as "I," "his," "it," which are used instead of naming the person or thing itself | + | | | + | + | + | |
| Word choice | | | | | | | |
| Avoid little-known color names | | | + | | | | |
| Do not use difficult ideas such as metaphors | + | | + | + | + | + | + |
| Avoid clichéd words and expressions | | | + | + | | + | |
| Avoid useless or superfluous words that add nothing to the meaning | | | + | | | | |
| Avoid words like doesn't, wasn't, couldn't. Write the words out in full instead | + | | | | | | |
| Avoid repetitions | | + | | | | | |
| Avoid using abbreviations. If you have to use initials, explain them | | + | + | + | + | + | + |
| Be concrete, not abstract | | + | + | + | | + | |
| Use precise words | | | + | + | | | + |
| Only use technical terms with colleagues or professionals | | + | + | + | | + | + |
| Name the agent in a sentence | | + | | | | + | |
| Loanwords | | | | | | | |
| Avoid anglicisms in French | | + | + | | | | |
| Beware of false friends between languages | | + | + | | | | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Words and phrases** | | | | | | | |
| Do not use loanwords if there is a well-known equivalent in your primary language | | | + | + | + | + | + |
| Dealing with difficult words | | | | | | | |
| Difficult words need to be explained | + | | + | + | + | + | + |
| Use examples to explain things | + | | | | | + | |
| Long compound words are partially repeated after their first mention | | + | | | | + | |
| If there are a lot of difficult words, make a dictionary (list) of useful words. Place it at the end of the text | | | | | + | | + |

As can be seen from Table 2.1, most guides seem to agree on the general principles, meaning that the words should be short, easy-to-understand, and well-known to the reader. In terms of grammar, all guides agree that active verb forms are preferable to passive forms, and most agree that writers should prefer positive forms to negative ones. Most guides also warn against nominalizations.

Anaphora avoidance and unambiguity have received their own category because all guides warn against using synonyms, and most advise not to use anaphora. There are also rules against ambiguity in general, which include word and phrase levels: for example, one of the checkpoints in the Easy Finnish Indicator states, “Instructions meant for the reader are expressed clearly and unambiguously”.

In terms of word choice, most guides advise against metaphors, which are considered a difficult structure (similar to, for example, humor in Table 2.6). Concreteness and precision are also considered important. It is also strongly recommended to avoid or at least spell out abbreviations.

Most guides advise against using loanwords, with not one but two guides specifically warning against using anglicisms in French. It is also permissible to use loanwords if there is no equivalent in the primary language. Like the rule from the previous category, which suggests that terminology can be used only if the audience is familiar with it (“Only use technical terms with colleagues or professionals”), the general principle of familiarity overrides specific rules if necessary.

When dealing with difficult words, almost all guides recommend providing explanations. Some advise using examples to explain things; others even recommend making a glossary at the end of a text. Meanwhile, some guides permit only partially repeating long compound words after the first mention. In both cases, it is about names: for example, the Finnish guide permits referring to the Ministry of Agriculture and Forestry as “Ministry” after the first mention (Ruel et al., 2018),

and the EC guide suggests that after naming the Committee on the Procurement of Language Style Guides with its full name, it can be referred to as just "committee" in the next sentence (European Commission, 2015). So, here the principle of brevity is allowed to override the principle of unambiguity.

### 2.3.2 Numbers and formatting

**Table 2.2.** Rules for numbers.

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Numbers** | | | | | | | |
| Prefer the numerical form of numbers | + | | + | | | + | + |
| Write the dates in full | + | | + | | + | | + |
| Give the equivalences of the metric system in (imperial) measurements | | | + | | | | |
| Avoid fractions | | | + | | | | |
| Use orders of magnitude or general words such as many, most, half, few, etc., instead of citing percentages, numbers, and statistics | | | + | | + | | + |
| Place the event in time with simple benchmarks known to the recipient public | | | + | | + | | |
| Avoid the use of Roman numerals | + | | + | + | + | | + |
| Write down the measure of length and weight in full | | | | | + | | |
| Use spaces or hyphens when writing phone numbers | | | | | + | | + |
| Be careful with numbers like "7th" meeting. It might be hard for some people to understand. You could write something like "The meeting we have had 6 times before" | + | | | | | | |

Most guides have a few rules on how to format numeric information, such as numbers, dates, measurements, or phone numbers. Most guides seem to agree that the Roman numerals should be avoided, with some recommending using only plain Arabic numbers (for example, the Dutch guide advises avoiding Roman numerals, letters, and combinations of numbers and letters when numbering lists). The majority of guides also recommend writing dates in full and preferring the numerical form of numbers. In the guides that lean more towards Easy Language, there is also advice on using orders of magnitude instead of precise numbers and

time benchmarks instead of precise dates (in the French guide, these rules are marked as being aimed at people with disabilities).

**Table 2.3.** Formatting rules.

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Formatting** | | | | | | | |
| Never use footnotes | + | | + | | + | | |
| Avoid all special characters where possible, like &, <, § or # | + | | + | | | | + |
| Use simple punctuation | | | + | | + | | |
| Avoid the use of parentheses, asterisks, hyphens in the text, ellipsis, slashes | | | + | | | | |
| Do not write entire words in capitals | + | | | | | | |
| Use the correct bullet point for a list or summary | | | | + | | | |
| Place explanations directly in the text or in boxes, rather than in hyphens or parentheses | | | + | | | | |
| Contrast "good" and "bad" solutions using drawings that help to better integrate the information, by caricaturing and using the colors of the traffic lights, red/orange/green, which convey meaning | | | + | | | | |
| Highlight the important elements. | | | + | | + | | + |
| Do not use different techniques to highlight important information in the same text. Stick to only one technique | | | | | + | | |
| Highlight negatives in bold | | | | | | | + |
| Avoid hyphenation at the end of a line | + | | + | | + | | |
| Separate long words with a hyphen | | | | | | | + |

The guides that I use actually have a lot more information on formatting and text layout than mentioned in Table 2.3, but as I am focusing mostly on the text, I consider only a few formatting rules. Most guides advise against making stumbling blocks for the reader's eye, such as special characters or hyphenation at the end of the line. Separating long words with a hyphen is German-specific advice, although some works suggest using a mediopoint (Bredel and Maaß, 2017): such segmentation has been proven to be beneficial for readers (Deilen, 2022). Even

though some guides suggest making a glossary of difficult words at the end of the text (see Table 2.1), multiple guides are against footnotes, probably also due to the fact that the reader's eye will stumble on it and the reader's attention will be distracted. Highlighting the important elements is also popular advice in the Easy Language-oriented guides.

### 2.3.3 Sentences

**Table 2.4**. Rules for sentences.

| **Rule** | **IE** | **EC** | **Fr** | **Nl** | **Ru** | **Fi** | **De** |
|---|---|---|---|---|---|---|---|
| **Sentences** | | | | | | | |
| Length and appearance | | | | | | | |
| Always keep your sentences short | + | + | + | | + | + | + |
| Always start a new sentence on a new line | + | | + | | + | | |
| Where possible, 1 sentence should fit on 1 line | + | | + | | + | | |
| Separate long sentences into shorter ones | | | | | + | | + |
| Prefer sentences of 7 to 12 words or 30 to 60 characters | | | + | | | | |
| Organization of information | | | | | | | |
| Do not bury important information in the middle of the sentence | | + | | + | | | |
| In sentences, the main clause and the subordinate clause are in a logical order in terms of how the text progresses | | | | | | + | |
| Do not clutter your document with redundant expressions like 'as is well known,' 'it is generally accepted that,' 'in my personal opinion,' 'and so on and so forth,' 'both from the point of view of A and from the point of view of B,' etc. | | + | | | | | |
| Try to give your sentences strong endings — that is the part that readers will remember | | + | | | | | |
| Interrupt the sentence structure with additional information as little as possible (object must be closer to the verb) | | | | + | | + | |
| The text should contain no structures in which a subordinate clause is in the middle of | | | | | | + | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Sentences** | | | | | | | |
| the main clause, as opposed to the beginning or end | | | | | | | |
| The predicate is located at the beginning of the clause | | | | | | + | |
| Do not overstress sentential brackets (Satzklammer) | | | | + | | | |
| Do not use prepositional chains | | | | + | | | |
| Prefer simple grammatical structures: subject, verb, complement | | | + | | | + | |
| Formulate simple sentences | | | + | | | + | + |
| Use only one subject if possible | | | | | + | | |
| Formulate your sentences in such a way that one sentence conveys one thought | | | | | + | + | + |
| In sentences, the clauses are bound together by conjunctions, (because, but, as) so that the relationship between the two items is clear | | | | | | + | |
| Avoid questions in the text. | | | | | | | + |
| Grammar and syntax | | | | | | | |
| Avoid complements before the verb | | | + | | | | |
| The text should not contain many language structures rated difficult | | | | | | + | |
| The text contains no non-finite clauses phrases or other non-finite structures | | | | | | + | |
| Nouns should not have multiple modifiers in one clause | | | | | | + | |
| The text contains mainly clauses that are structured on the basis of a finite verb | | | | | | + | |
| Sentences should not contain complex negative relationships | | | | | | + | |
| Avoid subordinate clauses | | | | | | | + |
| The use of tenses and time expressions in the text is consistent | | | | | | + | |

Unsurprisingly, most guides agree that sentences in simple texts should be short. Easy Language-oriented guides also suggest starting each sentence on a new line and keeping them one line long.

There seems to be no rule on organizing the information within a sentence that would be common to most languages. The experts seem to agree that sentences should be simple and one sentence should convey one thought. Some guides also seem to agree that important information should not be hidden in the middle of the sentence and/or that the relaying of important information should not be interrupted. The Finnish guide also highlights the need for logical presentation of information within a sentence: putting the clauses in a logical order and binding them together with conjunctions.

In the Grammar and Syntax category, Finnish-specific rules prevail, which is understandable given the complexity of Finnish grammar.

### 2.3.4 Text structure

**Table 2.5.** Rules for structuring the text.

| **Rule** | **IE** | **EC** | **Fr** | **Nl** | **Ru** | **Fi** | **De** |
|---|---|---|---|---|---|---|---|
| **Structure** | | | | | | | |
| Document structure | | | | | | | |
| Highlight the plan clearly | | | + | | | | |
| Structure the text with headings and subheadings | | | + | | | + | + |
| Use headings that are clear and easy to understand | + | + | + | + | + | + | |
| Try not to use too many layers of subtitles or bullet points | + | | + | + | + | | |
| Summary | | + | + | + | | | |
| Conclusion | | + | + | + | | | |
| Table of contents | | + | + | | | + | |
| It can be very useful to formulate the subheadings in question form | | | + | + | | | + |
| Add a title or introductory sentence to a list or summary | | | + | + | | | |
| Keep the number of elements in a list or summary limited | | | | + | | + | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Structure** | | | | | | | |
| Standardize the presentation and wording of titles as much as possible | | | + | | | | |
| The content matches the heading | | | | | | + | |
| Avoid references | | | | | | | + |
| Organization of information | | | | | | | |
| Always put your information in an order that is easy to understand and follow | + | | + | | | + | |
| Group all information about the same topic together | + | | + | + | | | + |
| It is OK to repeat important information | + | | | + | | | |
| It is OK to explain difficult words more than once | + | | | | | | |
| Organize the text into different logical parts | | | + | | | | |
| Repeat the most important information at the end | | | + | | | | |
| Separate main issues from side issues | | | | + | | | |
| Immediately or after a short step-up, make it clear what your intention is | | | + | + | | + | |
| Place important information first, then secondary and specialized information, and finally, conditions and exceptions | | | + | | | | |
| Link words, sentences, and paragraphs using transitions and relationship markers such as: first, then, because, in conclusion, etc. | | | + | | | + | |
| When a situation precedes an action, respect the order of appearance | | | + | | | | |
| There are no gaps in the content of the text. At every point of the text, the reader receives sufficient information to understand the text | | | | | | + | |
| The text is not too concise; neither is too much information packed into one clause | | | | | | + | |
| The text does not refer to what was said earlier in a way that assumes that the reader | | | | | | + | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Structure** | | | | | | | |
| will remember or find an earlier part in the text (see picture on page 3) | | | | | | | |
| Make sure the important information is easy to find. | + | | + | + | | + | |
| Avoid distracting elements, for example, several dates or numbers in the same paragraph. | | | + | | | | |
| Add context to new information and concepts | | | + | | | | |
| Avoid long definitions | | | + | | | | |
| Choice of information | | | | | | | |
| Always make sure you give people all the information they need | + | | | | | | |
| Only give the important information | + | + | + | + | | + | |
| Imagine what questions they might ask and make sure the document answers them. | | + | + | + | | | |
| Present information adapted to the target audience, according to its knowledge of the subject, its logic, its culture and its experience | | | + | | | + | |
| Mention accessible documents, sites, or videos to complement the information | | | + | | | | |
| Paragraphs | | | | | | | |
| Start a new paragraph when you start a new element in your text | | | | + | | | |
| Formulate the main idea of a paragraph in a clear topic sentence | | | + | + | | | |
| Start each new paragraph on a new line. | | | | + | | | |
| Keep an eye on the length of the paragraphs | | | + | + | | | |
| Limit the number of messages: present only one idea per sentence and one main idea per paragraph | | | + | | | | |
| Text quality | | | | | | | |
| The text contains no factual errors | | | | | | + | |
| The text follows the spelling rules of standard language | | | | | | + | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Structure** | | | | | | | |
| The text follows the typical characteristics of the type it represents | | | | | | + | |

In terms of document structure, most guides mention the need for clear and easy-to-understand headings and subheadings. There are also specific rules on nested structures such as bullet points, subheadings, and lists, with multiple guides (French and Dutch) agreeing on the maximum of 3 levels. According to "In duidelijk Nederlands," the maximum number of elements in a list or summary should be limited to 7. Interestingly, formulating subheadings in question form is considered preferable even in the German guide, which also advises against questions in the text (see Table 2.4).

The most popular advice on organizing the information within a document appears to be making the important information easy to find and grouping together information on the same topic. The importance of ordering information in a way that is easy to follow and making logical links in the text is highlighted again, just like on the sentence level, but this time by more guides. Some guides also mention making the text's intention clear immediately.

In terms of the choice of information, again, the most important rule seems to be only giving information that is important to the target audience. The experts advise thinking about the questions that the readers might ask and adapting the information to them.

There are a few rules concerning paragraphs specifically. Not all guidelines have them, but the ones that do suggest monitoring the length of the paragraphs (4-5 lines in the French guide, 6-7 sentences in the Dutch guide) and formulating the main idea in the beginning of the paragraph, just as the text's intention should be formulated immediately.

Finally, the Easy Finnish Indication has a set of rules specifically on the correctness of the text. These rules do seem important, and it is unclear why other sources do not mention them; perhaps it seemed too obvious to mention.

### 2.3.5 Tone of the text

**Table 2.6.** Rules on the tone of the text.

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Tone** | | | | | | | |
| Pronouns | | | | | | | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Tone** | | | | | | | |
| Use the pronoun Sie | | | | | | | + |
| Address the listener, conversation partner, or reader directly with the personal pronoun “u” or “je” | | | | + | | | |
| Use of humor and figurative language | | | | | | | |
| Avoid sophisticated and opaque forms of imagery and puns | | | + | + | | | |
| Be careful with ironic and other humorous elements in bad or unpleasant news | | | | + | | | |
| Be careful with expressive elements from the informal atmosphere | | | | + | | | |
| Addressing the reader | | | | | | | |
| Address the listener, conversation partner, or reader directly | + | | | + | + | + | + |
| Do not present instructions for the reader in the passive voice | | | | | | + | |
| The text linguistically distinguishes between what is mandatory for the reader and what is possible or recommended | | | | | | + | |
| Things are described on a general level, or the text addresses the reader indirectly when addressing them directly seems unnatural <…> | | | | | | + | |
| Use empathetic language that promotes closeness | | | + | | | | |
| Choose a positive or neutral tone | | | + | | | | |
| Choose a courteous and respectful tone | | | + | | | | |
| Avoid the authoritarian, threatening, moralizing, or accusatory tone | | | + | | | | |
| Avoid the use of the third person | | | + | | | | |
| The tone of the text is appropriate for the situation | | | | | | + | |
| The text does not underestimate the reader | | | | | | + | |

| Rule | IE | EC | Fr | Nl | Ru | Fi | De |
|---|---|---|---|---|---|---|---|
| **Tone** | | | | | | | |
| The reader is not presented as passive or as needing help too often; the reader is also an active party | | | | | | + | + |

Not all guides have specific rules about the tone of the simple texts. There is language-specific advice on pronoun usage for Dutch and German. Strangely, there are no such rules for Russian, which also has polite and informal forms of "you"; however, the examples in the Russian guide address the readers with "вы", which is simultaneously the polite "you" and the plural "you." This corresponds with the Russian tradition of formal and semi-formal written communication.

Similar to the rules on metaphor usage in Table 2.1, the experts advise against complex forms of expressive language, especially humor (irony, puns, etc.).

In terms of addressing the reader, most experts agree that the reader should be addressed directly and presented as an active party, with some guides specifically warning against the use of the third person.

## 2.4 General Principles of Simple Language

In the previous section, I touched upon basic principles of Simple Language, such as the principle of familiarity. Here, I will describe the general principles that encompass all or most of the guides examined. These principles are rather abstract, and a lot of the examined rules combine some of them.

**Commonness and familiarity:** some rules point to the fact that the words and phrases should be familiar to the reader and/or frequently used in the language. Uncommon words should only be used if they are well-known to the reader (see the "Word choice" section in Table 2.1). As stated under the aforementioned Table, this principle can override specific rules: for example, the use of loanwords is not advised but permissible if the word is already common in the language.

**Brevity:** words, sentences, and paragraphs are to be kept short. So should bullet points, lists, and texts in general. This principle can and should be overridden in favor of providing explanations.

**Explanations:** in most guides, there are rules that pertain to this principle explicitly: see, for instance, the section "Dealing with difficult words" in Table 2.1. This principle also covers spelling out abbreviations and writing dates and other measurements in full. It can override the principle of brevity, but can also be

overridden by it: for example, one of the rules in the aforementioned section permits using shorter entity names if the full name has just been spelled out.

**Logic and consistency:** there are multiple rules requiring consistency in word choice (including a very popular rule from Table 2.1 on using the same word for the same thing throughout the document), grammar (conjugating the verbs in the same tense), and document layout (in Table 2.3: stick to only one style of highlighting important information).

The principle of logic governs many of the structural rules (see Table 2.5). Most of these rules, such as organizing the text into different logical parts, linking parts of the text with transitions and relationship markers, and making sure the content matches the heading, can be viewed simply as general principles of any well-written informative text.

**Unambiguity and precision:** again, the majority of guides recommend avoiding ambiguity. Some advise against using polysemic words, while others recommend avoiding ambiguity across multiple phrases. Multiple guides also emphasize that ambiguity should be avoided specifically when giving instructions to the readers: for example, they recommend using imperative forms for instructions. The words and expressions also must be concrete and precise: for example, abstract lexis and metaphors are to be avoided.

**Focusing on and respecting the reader:** the Easy and Plain Language guides often address the appropriate tone for addressing the readers. The rules emphasize the importance of including the readers in the conversation: addressing them directly, using empathetic language, and presenting them as active parties. The authors of simple texts are also encouraged to choose the information with the readers' needs in mind.

Some rules explicitly remind the authors to respect the readers and choose a tone that is appropriate for the situation. Perhaps the rules on text quality (see Table 2.5) can also be attributed to this principle.

**Simplicity and clarity:** finally, most of the rules from the guides can be attributed to general rules of simplicity and clarity. Some rules merely state that the words and sentence structures should be simple and easy to understand, perhaps suggesting that the experts rely on their knowledge or lived experience. There is certainly a set of language features, such as active voice, present tense, and positive forms, that are universally considered simple. There is also a set of features, such as the conditional, loanwords that have not been firmly established in the language, and prepositional chains, that are universally considered difficult, and therefore should be avoided. Some rules of simplicity can be considered typographical: for

example, most of the rules in Table 2.3, or the rules about putting each sentence on its own line and making sure one sentence fits one line (see Table 2.4).

Many structural rules combine the principles of logic and clarity. The majority of such rules, for example giving the text a clear structure (summary, table of contents, conclusion, and the like), can be applied to any informative text. One could argue that some rules of sentence building, namely the rules of order (not burying the important information in the middle of the sentence, interrupting the sentence structure as little as possible, choosing simple grammatical structures, avoiding sentential brackets, etc.), also pertain to this general principle.

No rules completely override the principle of simplicity. When a difficult element must be introduced, there should either be a workaround - for example, providing a dictionary of difficult words - or the difficulty should be considered normal by the audience (such as specialized jargon in texts for professionals). At the same time, this principle must work in harmony with the rest. For example, a text written using only words from BASIC English (Ogden, 1940) may not actually be simple if some concepts are explained in a way that is uncommon for the readers' primary language (thus violating the principle of commonness and familiarity) due to the lack of vocabulary.

It seems that if the text follows all these general principles thoroughly, it can be considered written in Plain Language or, at the very least, an easy-to-understand text for experts. It is the degree of simplicity that determines where the text will be on the scale from Easy to Plain Language.

Finally, producing a text in Easy or Plain Language is a creative process. While the combinations of these general principles should aid the author in creating a simple text, the author can decide to violate some of these principles based on his or her expertise and creative vision. However, studying these principles and the specific rules is important for understanding what exactly makes a text simple and why.

## 2.5 Writing Strategies in Simple Texts

The previous sections covered the linguistic properties of Simple Language according to Easy/Plain Language guidelines. In this section, I will touch upon the strategies used by authors of simple texts "in real life." To do that, I will relay the findings of Paper III, which dealt with the linguistic features of texts from Russian textbooks for young speakers in various language environments.

In this paper, my co-authors and I examined several texts from textbooks aimed at native, non-native, and bilingual young (aged 7-11 years) speakers of Russian. The difference between the latter two groups can be understood as follows: bilingual speakers exist in an unbalanced bilingual environment with Russian being their

weaker language, and non-native speakers study Russian as a foreign language outside the Russian language environment. The main goal of the study was to find whether there are any specific simplification strategies in educational texts for children with different language backgrounds and levels of Russian language proficiency.

We used the TIRTEC[12] corpus as the dataset for this study (Laposhina, Veselovskaya, and Krivenko, 2019). The dataset statistics can be seen in Table 2.7. In the table, "R-native" denotes texts for native Russian speakers, "R-bilingual" - texts for bilingual speakers, and "R-foreign" - texts for foreign speakers.

**Table 2.7.** [Table 2 from Paper III] Dataset statistics.

| | **R-foreign** | **R-bilingual** | **R-native** |
|---|---|---|---|
| **Collection size** | | | |
| Number of texts | 1100 | 1100 | 1100 |
| Number of tokens | 39955 | 58964 | 31670 |
| Vocabulary size (number of unique tokens) | 8846 | 12760 | 10919 |
| **Text source** | | | |
| Simple fragment of authentic text | 170 | 205 | 727 |
| Fragment of authentic text adapted by textbook authors | 61 | 41 | 30 |
| Texts written specifically for this textbook | 869 | 854 | 343 |
| **Basic text characteristics** | | | |
| Mean sentence length (words) | 5.84 | 7.3 | 7.56 |
| Mean word length | 4.67 | 4.86 | 5.14 |
| Average number of punctuations per sentence | 0.73 | 0.81 | 1.02 |

[12] https://digitalpushkin.tilda.ws/tirtec

As can be seen, there are both inherently simple and simplified (adapted) Russian texts in the collection. It can also be seen that, although the R-native texts have the fewest tokens, they are the most diverse in vocabulary: the ratio of unique words to all words is the highest in this category. This indicates that the authors of the non-native textbooks could have been limiting the usage of rare and "difficult" words.

In order to study the linguistic properties of our texts, we extracted a total of 95 quantitative features from our data. There were length-based features (such as average sentence and word length), readability scores, lexical features (mostly coverage by various vocabulary lists, such as the lexical minima for different CEFR grade levels of Russian), and morphosyntactic features (percentage of nouns, words in the genitive case, etc.). Our independent variable was the text's class: R-native, R-bilingual, or R-foreign. We looked at the correlations between different features and fit multiple logistic regression models in an attempt to see which features would influence the chosen class the most.

As can be seen in Table 2.8, we did not find any strong correlations. However, we did see some distinctions: for example, although R-bilingual texts have a lot of unique words (i.e. absolute number of non-repeating lemmas in one text), these words seem to be frequently used in Russian. It can also be argued that negative correlation with type-token ratio and lexical density indicates that texts of this class do not in fact have a lot of lexical variety, and the absolute numbers of unique words in these texts are higher simply because the texts tend to be longer. Texts aimed at learners of Russian as a foreign language have more vocabulary from the established lexical minima. They also seem to avoid verbs in past tense, which corresponds to some of the Simple Language guidelines. On the contrary, texts for native speakers have the fewest "certified easy" words. They do, however, perform very well on some readability scores. They also tend to have higher type-token ratios, which indicates higher variety in the vocabulary.

**Table 2.8.** [Table 3 from Paper III] Kendall's tau correlations.

| **Domain** | **Most significant features** | **Kendall's tau** |
|---|---|---|
| R-foreign | Relative numbers of verbs in past tense | -0.28 |
| | Percentage of A1 vocabulary | 0.28 |
| | Percentage of A2 vocabulary | 0.26 |
| | Coleman's readability formula | -0.25 |
| | Percentage of B1 vocabulary | 0.25 |
| | Relative numbers of verbs in perfective aspect | -0.24 |
| R-bilingual | Number of unique words | 0.19 |

| Domain | Most significant features | Kendall's tau |
|---|---|---|
| | Number of words | 0.19 |
| | Text coverage by 5 000 most frequent Russian words list | 0.18 |
| | Relative amount of nouns | -0.13 |
| | Lexical density | -0.12 |
| | TTR | -0.12 |
| R-native | Percentage of A1 vocabulary | -0.39 |
| | Percentage of B2 vocabulary | -0.39 |
| | Percentage of B1 vocabulary | -0.38 |
| | TTR | 0.33 |
| | Coleman's readability formula | 0.26 |
| | ARI readability formula | 0.25 |

We employed multi-class and binary (in one-vs-the-rest fashion) logistic regression models in order to further study the influence of the linguistic features on the text's class. In the one-vs-the-rest setting, R-native texts proved to be the most distinguishable from the others (F1-score of 0.78), while R-bilingual texts were the hardest to classify (F1-score of 0.68).

The error analysis of the three binary classifiers showed that, for all types of errors, the median value of the percentage of words from lexical minima turned out to be closer to the median value not of its correct category, but of the one determined by the model. For instance, R-native texts that were marked R-bilingual by the model tended to contain more vocabulary from the CEFR-graded lexical minima than average R-native texts. In R-native texts marked as R-foreign, these numbers were even higher. This can indicate that lexical differences were one of the factors that confused the model. Readability proved to be among such factors as well. Some grammatical features, such as relative numbers of adjectives, nouns, verbs, and adverbs among all words, also influenced the wrong decisions of the model. For example, in R-native texts, the relative number of adjectives is quite high on average. However, in R-native texts that were wrongly identified as R-bilingual this number is lower, and in R-native texts marked as R-foreign there were almost no adjectives at all. Finally, it is worth noting that the model made more errors on texts from certain textbooks, which may indicate that these texts did not correspond to the proclaimed target audience. This is especially true for the most diffuse category, R-bilingual.

From this study, it can be seen that, on the one hand, some of the rules found in the Simple Language guides are clearly used by authors of "real" simple texts and simplified adaptations. Because this study focused on quantitative features, it is unclear whether the structural principles such as providing explanations, logic and consistency, and unambiguity were always adopted. However, it can be seen that texts that are supposed to be relatively simpler adhere to the principles of brevity, commonness/familiarity, and general simplicity. For example, a text from a Russian as a foreign language textbook is generally supposed to be simpler than an educational text for native speakers, and our study has shown that R-foreign texts do in fact have shorter words and sentences (see Table 2.7), vocabulary that is supposed to be familiar to the reader from the established lexical minima, and fewer "difficult" words such as verbs in the past tense and adjectives. On the other hand, the rules that the authors of these texts apply are clearly domain-specific: for instance, the authors of educational texts for non-native audiences seem to be more concerned with the vocabulary adhering to the established lexical minima than are the authors of texts for native speakers.

In Paper VII, my co-author and I compared Easy Finnish news articles to their corresponding standard Finnish versions. My co-author, who studied how the linguistic properties of Easy Finnish news articles correspond to the Selkomittari guide, found out that the principles of brevity and simplicity are implemented in many of the texts. This includes more general rules such as "clauses and sentences are mainly short" and "sentence structures are simple," as well as more specific ones such as "the text contains no words that have several different elements, such as derivative affixes, inflectional suffixes, and clitics" (these rules are also discussed in more detail in the end of Chapter 3). It should be noted that the editing team of the Yle Easy Finnish News uses their own (privately developed) guidelines and also artistic freedom in making Easy Finnish content rather than external guides such as Selkomittari (P. Seppä, personal conversation, March 2023).

To conclude, it is evident that the linguistic properties of Simple Language described in the Easy/Plain Language guides are actually present in simple and simplified texts across domains and languages. Given that there are still no popular and widely used Simple Russian guidelines, and Selkomittari is a guide for evaluation rather than for writing, it can be presumed that the authors of the simple texts do not follow strict rules but instead seem to have some kind of linguistic intuition that guides them in simplification. (Although sometimes they do follow specific instructions, such as the authors of R-foreign textbooks using established lexical minima). In other words, although the general principle of simplicity is hard to define in precise terms, many of the same simplification strategies described in various Easy/Plain Language guides are still followed by different authors across domains and languages. However, at the same time, the authors of simple texts that

seemingly belong to the same genre (such as educational texts for young learners) may use different writing strategies depending on the audience.

## 2.6 Simplification Guidelines and Automatic Simplification

The next chapters of this thesis will deal with automatic text simplification. In Section 2.2 I mentioned that language technologies help facilitate accessible communication in multiple ways. However, does this mean that the rules of Easy and Plain Language govern automatic text simplification?

The answer to this question would be different for every model, because it depends on the data that the model was trained and fine-tuned on. The model can only learn the behaviors that are present in the data. Therefore, its simplification abilities will depend on the degree of simplification in the training corpus.

Most of the time, parallel corpora for TS are constructed from available data such as simplified news, Simple Wikis, or adapted books. It is assumed that the authors of these texts followed some kind of simplification rules, but these rules are rarely explicitly laid out. It is remarkably rare for a training corpus to be constructed from scratch just for the purpose of training simplification models. Most of the time, only the validation/test sets are made by humans, such as the RSSE validation dataset in Sakhovskiy et al. (2021). An exception would be the Amazon Mechanical Turk corpus, where 2350 sentences have been simplified by 8 different Turk workers, thus creating a big enough dataset to be used not only for testing and validation, but also for fine-tuning. However, it would be very difficult to find a parallel dataset suitable for simplification in which the sentences were simplified according to Easy/Plain Language guidelines. For instance, in the two aforementioned datasets, the crowd-source platform workers were instructed to simplify texts, but were not given any elaborate rules for simplification besides meaning preservation and, in the Turk's case, content preservation and not splitting sentences. It is safe to assume that most of the simplifications in these corpora were not directly influenced by any established Easy/Plain Language guides.

It should be noted that strictly following a simplification guide similar to those covered in this chapter would be difficult both in a rule-based simplification system and in a data-driven one. A rule-based approach will not work for complex discourse-level rules such as the rules of tone in Table 2.6. A data-driven system would require a large variety of texts simplified by professionals who are familiar with the task of text adaptation and can follow the guidelines carefully. Ideally, all editors should agree on the meaning of the more abstract rules, such as "describing things on a general level," before starting the adaptation work. That is why, even when simplification datasets are created by hand, the simplification rules that the editors receive are generally non-specific.

In conclusion, we can assume that the simplification models will adopt the general principle of simplicity. It may be possible to infer some simplification principles that the model has adopted, such as using more common vocabulary or splitting long sentences, but in general, a data-driven model will probably not follow any particular simplification rule consistently. However, a good model will demonstrate some degree of simplification measurable by automated evaluation metrics and visible to human assessors. And, since at present following specific simplification rules is virtually impossible in modeling, adhering to the general principle of simplicity and clarity is the necessary and sufficient condition for today's simplification models. This approach also allows us to account for domain-specific simplification. Because simplification rules can vary a lot depending on the audience, it is unfeasible to create a dataset for every simplification domain. However, it is possible to make a simplification model on the basis of the existing texts for this audience and expect it to learn the strategies that the writers in this domain generally adopt. Finally, because, as discussed above, the Easy/Plain Language guides also seem to be derived "from practice" (i.e. from the existing simple and simplified texts that were received well by their readers), it can be assumed that a native speaker just following her implicit idea of what simplicity means in a text will produce a text that follows at least some of the guidelines; therefore, a model trained on texts simplified by native speakers, especially by professionals (journalists or book editors), will learn to mimic that implicit knowledge.

# 3 Data

## 3.1 Data sources for text simplification

This chapter will cover the development of two parallel data sources for text simplification: RuAdapt, the Parallel Russian-Simple Russian dataset (Paper II), including the RuAdapt Word Lists (Paper IV), and the Parallel Corpora of Finnish and Easy-to-read Finnish (Papers VII and VIII). In this chapter I will outline the process of fulfilling the first research objective of creating new simplification-specific data sources for languages that lack such datasets and making these resources accessible at least for academic use. I will also detail the various strategies used to create the aforementioned dataset in order to cover the second research question of this thesis (RQ2).

Text simplification datasets are typically monolingual parallel datasets in which each "regular" segment, be it a text, a paragraph, or a sentence, has a simpler equivalent that conveys roughly the same information. These datasets structurally resemble machine translation datasets.

Until recently, most text simplification datasets were made for English. For example, one of the first datasets made specifically for automatic sentence simplification was constructed on the basis of Simple English Wikipedia and English Wikipedia (Coster and Kauchak, 2011). A notable exception was the Newsela dataset, available in English and Spanish (Xu et al., 2015). Unfortunately, although the Newsela data is high in quality, the dataset is only available on request, which has made it difficult for the research community to use it.

However, the last decade saw a spike in the development of non-English simplification datasets, including both European languages such as French (Gala et al., 2020), German (Stodden et al., 2023), Spanish (Moreno-Sandoval et al., 2024), and Italian (Tonelli et al., 2016), and non-European languages, such as Japanese (Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018) and Arabic (Khallaf et al., 2022). Currently, there is also a multi-lingual text simplification benchmark available, made of open-access simplification data for many languages (Ryan et al., 2023).

There are other data sources besides parallel simplification corpora that can be used for studying simplification strategies or developing a simplification system. For example, simplification-like data can be mined from other monolingual parallel data, such as paraphrase or summarization datasets (these strategies were used for

data augmentation and knowledge transfer in Papers I and V). For some languages, there also exist word lists, such as synonym lists, that can be used for lexical simplification or for simplification evaluation. For example, for Russian, there exist so-called lexical minima: lists of words that are supposed to be known to a student at a certain level (CEFR grade) of Russian language acquisition. These lists are used, for example, as preparatory materials for the official Russian as a second language examination called TORFL (Test of Russian as a Foreign Language; rus. ТРКИ, государственное тестирование по русскому языку как иностранному языку) (see Андрюшина et al., 2019 as an example).

## 3.2 Parallel Russian-Simple Russian Datasets

### 3.2.1 The RuAdapt dataset

In Paper II we describe the process of creating a text simplification dataset for Russian which was given the name RuAdapt. Almost at the same time as RuAdapt's release, another parallel Russian-Simple Russian dataset was published to serve as a data source for a shared task on Russian text simplification (Sakhovskiy et al., 2021).

At the time of its creation, RuAdapt consisted of works of classical Russian literature and their simplified versions. The latter were simplified by professional editors and published by the Zlatoust publishing house[13], which has kindly granted us permission to use these works. Most of the original works were published long ago, which meant that the copyright restrictions have been lifted. There were, however, some works that did not make it to the open-access version of the dataset due to the original versions still being under copyright.

The Zlatoust books came into our possession in the same format as they were printed, meaning that we were given electronic books in PDF format with the same layout and graphics as the published versions. Therefore, the preprocessing of these books included converting PDFs to XML files with Apache Tika[14], cleaning the texts of noise, and converting them to plain text files. Cleaning included removing unnecessary line breaks, stress marks, and other noisy symbols with the help of a dedicated Python script. The original texts were downloaded from the web, cleaned in the same way, and stored in plain text files for further analysis and subsequent alignment.

It should be noted that the distribution of books from the Zlatoust book collection by CEFR grade level is uneven. Most of the texts are for level B1; however, for a large part of the collection, the CEFR level is very vaguely specified:

[13] http://zlat.spb.ru/

[14] https://tika.apache.org/

particularly, there are a few collections of novels for which the level is said to be between B1 and C1, as can be seen in Figure 3.1.

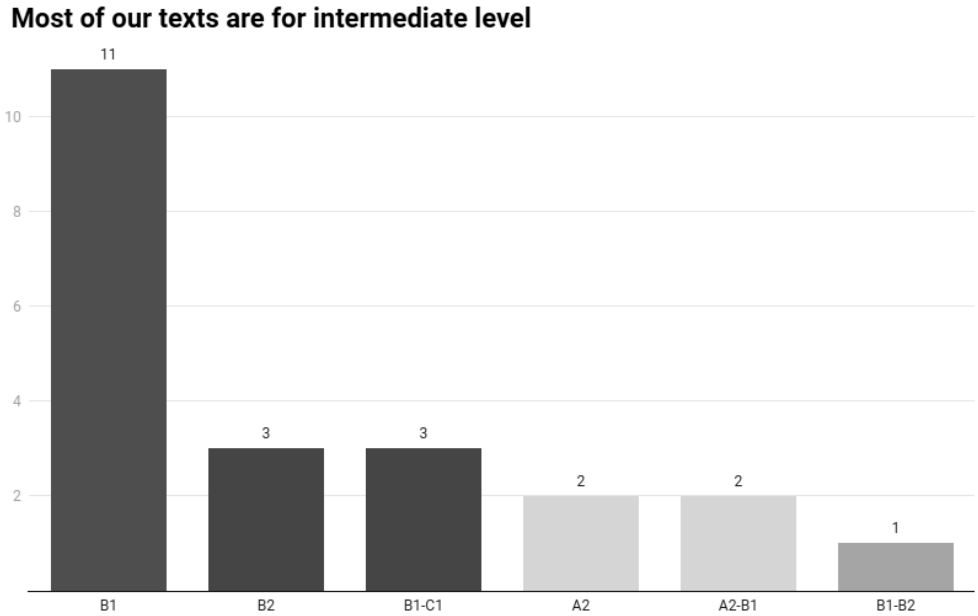


**Figure 3.1.** Books in the RuAdapt dataset, including the ones that could not be added to the published version of the dataset.

Some statistics from the original RuAdapt datasets can be found in Table 3.1. "TTR" stands for type-token ratio, which shows the degree of lexical variation. "FKGL" is the Flesch-Kinkaid grade level, a measure for calculating readability. We used the formula with constants adapted for Russian texts[15].

**Table 3.1.** [Table 1 from Paper II] Characteristics of adapted and original texts. Sent = sentence, par = paragraph. * - weighted average.

| **Metric** | **Original** | **Adapted** |
|---|---:|---:|
| Words | 885167 | 268409 |
| Unique words | 89318 | 32762 |
| Sentences | 69737 | 29003 |
| Par length / text* | 250.45 | 180.29 |
| Punctuation / sent* | 2.4 | 1.66 |
| Sentences / par* | 3.15 | 3.19 |
| Average TTR | 0.42 | 0.43 |

[15] https://github.com/infoculture/plainrussian

| Metric | Original | Adapted |
|---|---|---|
| Words / par* | 39.06 | 28.7 |
| Word length / text* | 5.08 | 4.89 |
| Average FKGL | 6.04 | 4.49 |

We also calculated the number of words in the texts that are supposed to be known to the reader at a certain level of language acquisition. This was done by comparing the vocabulary (unique lemmatized words) of the adapted and original books to lexical minima for learners of RaaFL (Russian as a Foreign Language). We considered named entities such as names and locations to be included in the readers' vocabulary, since the characters' names and locations in books are usually repeated numerous times. The results of the comparison can be seen in Table 3.2.

**Table 3.2.** [Table 2 from Paper II] Mean amount of known vocabulary (percentage of words in the lexical minimum).

| Level | Original | Adapted |
|---|---|---|
| A1 | 57.48 | 62.97 |
| A2 | 65.62 | 71.81 |
| B1 | 74.3 | 80.69 |
| B2 | 81.82 | 87.24 |
| C1 | 89.67 | 92.97 |

The first release of RuAdapt, described in Paper II, contained only a paragraph-aligned version of the dataset. The intention was to experiment with longer sequences, so simplifications would go beyond the sentence level. It was also a way to take into account sentence-level editing operations such as splitting, combination, and mixing sentences.

We used two aligners, Bleualign (Sennrich and Volk, 2010) and CATS-Align (Štajner et al., 2018). In Bleualign, the alignment is performed based on the modified BLEU score between source sentences translated into the target language and the original target language sentences. CATS-Align offers multiple options for alignment methods, but our choices were the language-agnostic character trigram similarity strategy that uses the log TF-IDF weighting and compares vectors with the cosine similarity, paragraph alignment level, and closest similarity alignment strategy. The aligners' performance was compared on a small manually aligned test set which consisted of 302 texts from textbooks for children-native Russian speakers (this data was never released due to copyright). Both aligners performed well, but CATS-Align showed better performance, with a 0.98 F1-score.

Currently, there are three subcorpora in RuAdapt. One is the adapted literature subcorpus described in Paper II. The second ("encyclopedic") one is comprised of entries from the Multimedia Linguistic and Cultural Dictionary created by researchers at Pushkin State Russian Language Institute[16]. The dictionary entries describe various elements of Russian culture as well as linguistic, geographical and other phenomena (Ростова, 2018). The third part consists of adapted Russian fairytales created by the Moscow State University Institute of Russian Language and Culture's linguocultural educational project[17]. All subcorpora have a paragraph-aligned and a sentence-aligned version made automatically with CATS-Align. Each pair of paragraphs and sentences has a cosine similarity score ranging from 0 to 1, where 1 means that the texts in the pair are identical. For downstream tasks such as text simplification, it is recommended to set a cosine similarity threshold, since pairs with low similarity may contain texts with different meanings.

The current dataset statistics can be found in Tables 3.3 and 3.4, and the dataset itself is accessible at https://github.com/Digital-Pushkin-Lab/RuAdapt.

**Table 3.3.** RuAdapt dataset statistics: paragraph-aligned version.

| Subcorpus | Texts | Original tokens | Adapted tokens | Paragraphs |
|---|---|---|---|---|
| Adapted literature | 93 | 620563 | 287358 | 7614 |
| Encyclopedic texts | 355 | 207252 | 144378 | 3517 |
| Fairytales | 9 | 7093 | 4652 | 134 |

**Table 3.4.** RuAdapt dataset statistics: sentence-aligned version.

| Subcorpus | Texts | Original tokens | Adapted tokens | Paragraphs |
|---|---|---|---|---|
| Adapted literature | 93 | 376432 | 285190 | 24232 |
| Encyclopedic texts | 355 | 161954 | 144250 | 10271 |
| Fairytales | 9 | 5775 | 4642 | 416 |

[16] https://ls.pushkininstitute.ru/

[17] https://www.skazki.irlc.msu.ru/

### 3.2.2 RuAdapt Word Lists

The current version of RuAdapt has been used in studying lexical simplification strategies in adapted texts and compiling a list of monolingual word alignments of complex Russian words and their simpler equivalents. These findings are described in detail in Paper IV.

Lexical simplification has attracted considerable attention in the past few years as a key task in facilitating reading comprehension for different target readerships (Saggion et al., 2022). Like other simplification subtasks, at present, it is dominated by English data: for example, in the recent TSAR-2022 shared task on multilingual lexical simplification for English, Portuguese, and Spanish, English lexical simplification quantitative results were noticeably better than those obtained for Spanish and Brazilian Portuguese (ibid.). However, new data for other languages continues to emerge, including datasets for lexical complexity evaluation, since correct evaluation of word complexity can be an important step in many lexical simplification pipelines (Абрамов and Иванов, 2022). An example of such a dataset for Russian is the Russian Synodal Bible-based dataset for lexical complexity evaluation, created specifically for predicting word-level complexity in Russian texts (ibid.).

In order to produce word alignments, we used the entire RuAdapt dataset to extract 15,156 sentence pairs with a cosine similarity score lower than 0.99 but higher than 0.31. These thresholds were chosen empirically to omit both identical pairs and pairs of sentences with different meanings.

Two different word aligners were used for the word alignment task: eflomal (Östling and Tiedemann, 2016) and awesome-align (Dou and Neubig, 2021). The former is based on a Bayesian model with Markov Chain Monte Carlo (MCMC) inference. It is a statistics-based approach that requires additional data to be trained on for tackling new tasks, so we used a set of Russian paraphrases for pre-training, taken from Opusparcus (Creutz, 2018) and ParaPhraser.ru (Gudkov et al., 2020). Awesome-align is a BERT-based aligner, which can also be fine-tuned, but this was not required, so we used it as is.

In Table 3.5, it can be seen that most of the alignments, as expected, were pairs of identical words. There were only 8403 unique pairs that both aligners identified. It should be noted that the texts had not been lemmatized before alignment, because some linguistic phenomena that we wanted to study, such as the use of archaic grammar forms, could have been lost during lemmatization.

**Table 3.5.** [Table 1 from Paper IV] Alignment statistics.

| **Statistic** | **eflomal** | **awesome-align** |
|---|---:|---:|
| All single word pairs | 188706 | 193778 |

| Statistic | eflomal | awesome-align |
|---|---|---|
| Pairs consisting of different words, cleaned from noise | 19687 | 22767 |
| Unique pairs | 14807 | 15989 |
| Unique pairs in common | | 8403 |

Altogether we obtained 22393 word pairs (considering that 8403 pairs were identical between the two aligners). In order to further clean the pairs from noise and remove non-synonymic pairs, we employed human editors.

A total of 18 editors worked on the RuAdapt word lists. All of the editors were students and/or specialists in teaching Russian as a foreign language from the Pushkin State Russian Language Institute. Editors were asked to give each word pair a score of 0, 1, or 2. A score of 2 is supposed to indicate that a pair can be included in the word alignment list; score of 0 indicates the opposite. Score 1 is given in cases of uncertainty that may be included in the list or preserved for future studies. It is important to note that we were aiming to evaluate the "usefulness" of the pair for the word list, not its alignment quality. When a pair is "useful," it means that it can provide some information on how the vocabulary changes during the simplification process.

Each word pair was evaluated by at least two editors. At the end of the editing process, a third editor re-evaluated the pairs that received a 2 from at least one editor. This evaluation resulted in the final list of 1409 word pairs. It appeared that 9.11% of awesome-align alignments and 8.08% of eflomal alignments received a score of 2. Since the scores do not reflect the alignment quality directly, they do not illustrate the aligners' efficiency, but instead give an idea of how many single-word alignments will end up being synonymous or semantically similar in some other way.

The list has 1,134 unique "complex" lemmas and 811 unique "simple" lemmas. We further inspected the pairs to find out if the "complex" words were actually more difficult and less frequently used than the "simple" words. To do so, we examined the CEFR grade levels of the words in pairs using the established lexical minima for Russian as a foreign language and their IPM values (instances per million words) using the frequency dictionary of modern Russian language (Ляшевская and Шаров, 2009). In the majority of cases, the "simpler" words in pairs were more frequent in the Russian language. As for the CEFR grade level, in 513 cases the "complex" word's CEFR level was higher, and in 545 cases the "complex" word was not present in the lexical minima while the "simple" word was. We noticed that most of the cases where the "simpler" word was not "simpler" (had the same CEFR level as the "complex" word) can be explained by the authors choosing a word whose derivative appears in the lexical minima for lower CEFR levels, so the reader is more

likely to guess its meaning. For example, the word сердиться [to be grumpy] is replaced by злиться [to be angry]; both verbs are B2 level, but the cognate adjective злой [angry ADJ] appears at the earlier A2 level. In cases where the simple word appears at a higher CEFR level than the complex word, the word choice might have been prompted by the desire to use an "international" synonym (e.g., расстояние [spacing, distance] -> дистанция [distance]). These cases are specific to the adapted texts for learners of Russian as a foreign language.

We have also observed the following domain-specific simplification strategies based on the word list: replacing an archaic grammatical form of a word with a modern one (e.g., простою [simple ADJ+GEN, archaic] -> простой [simple ADJ+GEN]) and replacing an obsolete word with a modern synonym (e.g., особливый [special, archaic] -> отдельный [special]). However, most of the list presents more universal types of lexical simplification, such as replacing a word with a more neutral and frequent analogue (e.g., умолять [to beg] -> просить [to please]), use of hypernyms (e.g., соловьи [nightingales] -> птицы [birds]), or the removal of subjective evaluation suffixes (e.g., деревенька [village+diminutive suffix] -> деревня [village]).

This research has proven that not a lot of text simplification actually happens just on the word level. In the future, it may be beneficial to explore phrase-level simplifications separately as well. The resulting list consists of 1409 word pairs and is available on GitHub: https://github.com/Digital-Pushkin-Lab/RuAdapt_Word_Lists. The list can be used for further studies of simplification strategies used in text adaptation for learners of Russian as a foreign language, or in refining simplification systems, since it represents the "expected" word replacements.

## 3.3 Parallel Corpora of Finnish and Easy-to-read Finnish

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 (Lindholm and Vanhatalo, 2021). 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.). However, until recently, there existed no parallel Finnish-Easy Finnish corpora. Hence my colleague and I created several parallel Finnish-Easy Finnish datasets on the basis of the Yle news corpora. We wanted to provide a resource for studying the simplification strategies used by simple language content providers as well as a database for future automatic Finnish text simplification studies.

### 3.3.1 Source data: the Yle news corpora

Yle is a Finnish state-owned national public broadcasting company. It provides content in multiple languages besides Finnish, including Easy Finnish. Yle Uutiset selkosuomeksi [Yle News in Easy Finnish] provides short (about 5 minute long) daily radio and TV broadcasts. The radio broadcasts are also published on Yle's website in text form. Typically, the editors of the Easy Finnish segment pick and translate the topics of the day that they deem most interesting and/or important for their target audience (P. Seppä, personal communication, March 3, 2023). Most "regular" news articles, if selected for Easy Finnish news, are translated and come on air within 24 hours; however, in rare cases, an older piece of news or an article from the Swedish Yle may be translated (ibid.).

All articles from the Standard Finnish Yle's official website, starting from 2011, have been preserved in the Language Bank of Finland (Kielipankki). Kielipankki also provides archives of Swedish and Easy Finnish Yle news. Most of the content that can be found on the actual web pages, including details like image ids and descriptions, thematic tags, and other meta information, is preserved in the archives. This has made the Yle archives on Kielipankki an obvious choice as a data source for a Finnish-Easy Finnish parallel corpus.

The following datasets were used as data sources for Papers VII and VIII:

- Yle Finnish News Archive, 2011-2018 (Yleisradio, 2017)
- Yle News Archive Easy-to-read Finnish, 2011-2018 (Yleisradio, 2019)
- Yle Finnish News Archive, 2019-2020 (Yleisradio, 2021a)
- Yle News Archive Easy-to-read Finnish, 2019-2020 (Yleisradio, 2021b)

It should be noted that the Easy Finnish news articles prior to September 2014 have no clear identifiers in the archives, i.e., for the earlier articles, there is no way to establish that they definitely came out on Yle Uutiset selkosuomeksi. Therefore, the parallel corpora span the period from September 2014 to December 2020.

There is also no definitive way to link the Standard and Easy Finnish articles. The "regular" sources of the Easy Finnish news are never specified explicitly. In order to create parallel data, we used methods of distributional semantics described in the next section. For sentence alignment, statistics-based and neural network-based approaches were used.

Because the source datasets' contents were retrieved from Yle's internal archives and deposited into Kielipankki with the help of Yle, the quality of the source data is very high. Therefore, little to no data cleaning was needed in making the parallel corpora.

### 3.3.2 Alignment

The Parallel Corpora of Finnish and Easy-to-Read Finnish are available in document-aligned and sentence-aligned formats. The document-aligned version is divided into two datasets (Dmitrieva et al., 2022, and Dmitrieva and Yleisradio, 2024a), because the work on this project originally started with the archives from 2019-2020. The first dataset (Dmitrieva et al., 2022) has been manually annotated by a human expert, and these annotations have aided in creating the other parallel datasets. The sentence-aligned version for articles from 2014-2020 is available as a single archive (Dmitrieva and Yleisradio, 2024b).

#### 3.3.2.1 Articles

Aligning the Easy Finnish articles with their Standard Finnish sources proved to be a challenging task. As mentioned above, there are no clear links between the articles, so we had to search for a match for any piece of Easy Finnish news among all Standard Finnish articles that came out before the Yle Uutiset Selkosuomeksi that day. To narrow the search down, only articles with matching subject tags were considered as potential pairs.

We used methods of vector semantics to find pairs of articles with the same subjects. Altogether, four different models were used:

1. LASER[18] (we used the laserembeddings library[19]),
2. LaBSE (Feng et al., 2022; we used the version from sentence-transformers[20]),
3. MPNet (Song et al., 2020; we used the version from sentence-transformers[21]),
4. DistilUSE (multilingual knowledge distilled version of multilingual Universal Sentence Encoder (Yang et al., 2020)), also from sentence-transformers[22].

For the first parallel dataset, the Parallel corpus of Finnish and Easy-to-read Finnish from the Yle news articles 2019-2020, we only used a simple matching method that involved getting the document vectors using the DistilUSE embeddings and finding pairs of documents with the highest cosine similarity. A document vector was derived by taking the average from the first 15 sentence vectors in the document. The 15 sentence limitation was enforced due to the lack of

[18] https://github.com/facebookresearch/LASER
[19] https://github.com/yannvgn/laserembeddings
[20] https://huggingface.co/sentence-transformers/LaBSE
[21] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2
[22] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

computing resources that we were experiencing at that time. This limitation was lifted in later experiments.

We had to establish a cosine similarity threshold to extract the meaningful pairs. As seen in Figure 3.2, 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. In order to assess whether the scores actually represented semantic similarity and, if so, decide on what the threshold should be, we performed human evaluation on articles with scores from 0.6 to 1.

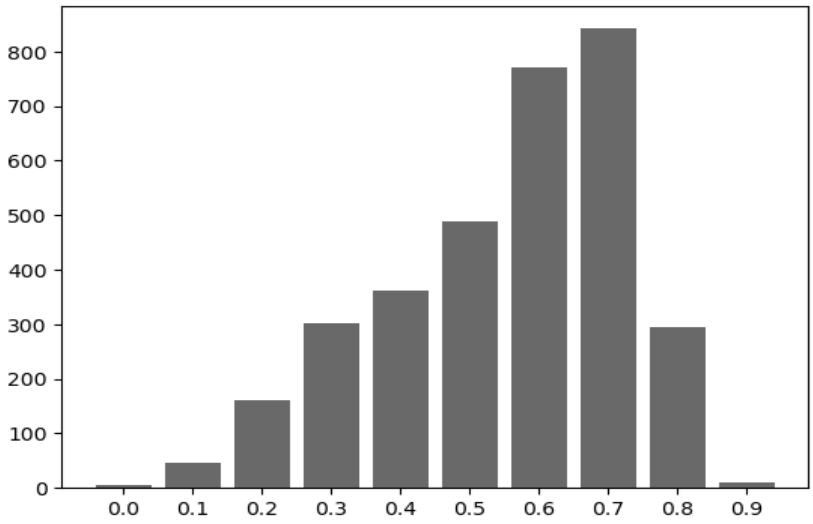


**Figure 3.2.** [Figure 1 from Paper VII] Distribution of cosine similarity scores across article pairs. X-axis is the approximated cosine similarity, y-axis is the number of pairs.

An expert was asked to evaluate each pair and give it one of three scores: “positive” – if the two 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 demonstrated on Figure 3.3, 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". It should be noted that we first performed automatic alignment, and only after that did the expert evaluate the aligned text pairs.

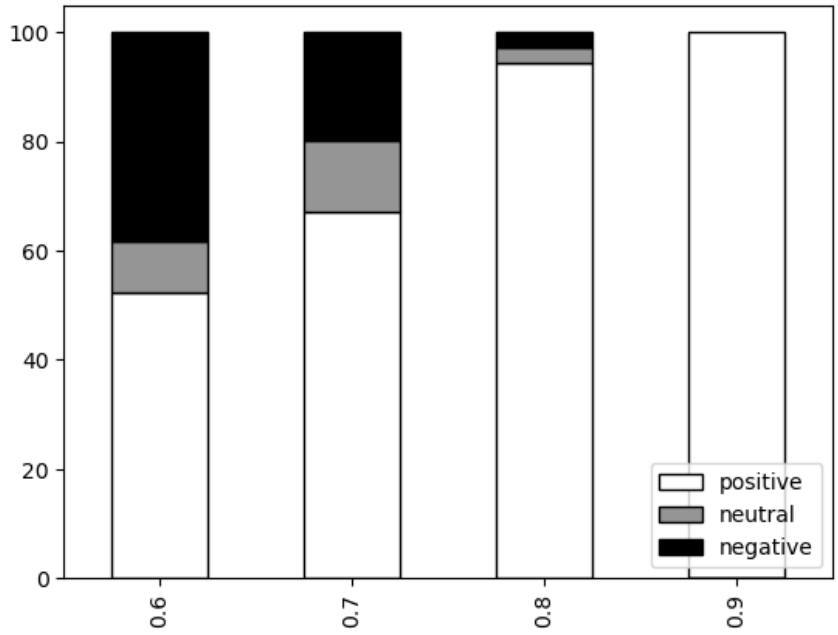


**Figure 3.3.** [Figure 2 from Paper VII] Percentages of labels given by the expert. X-axis is the approximated cosine similarity, y-axis is the percentage.

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

- The Easy Finnish article was about a completely different topic.
- The Easy Finnish article covered a similar topic but did not match exactly the original article for various reasons (e.g., time, location, different focus).
- The Easy Finnish article could not be mapped to one original article but compiled information from several original articles.

This small labeled dataset was later used in the creation of the Parallel corpus of Finnish and Easy-to-read Finnish from the Yle news articles 2014-2018. This time, more sophisticated strategies along with the simple vector matching approach were used for aligning the articles. Namely, we used a technique proposed in Thompson and Koehn (2020).

Firstly, we employed the script[23] provided in the Vecalign GitHub repository to obtain document embeddings for candidate generation. This method can be used

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

with different sentence embeddings, so we tried it with all four types of embeddings mentioned above. We set the K nearest neighbors to 5 and kept all other parameters default (such as J = 16 and γ = 20). We also experimented 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 used the dimensionality reduction technique proposed within the library[24]. For LASER, we used the PCA (principle component analysis) module from scikit-learn (Pedregosa et al., 2011).

Secondly, we utilized 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 did this for the documents aligned with the Vecalign method and, following the original paper, used Vecalign with LASER sentence embeddings. We modified the original re-scoring formula as follows:

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

Here, E and F are the Easy and Standard Finnish 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 did not divide by the total number of alignments, because the mismatch in sizes of source and target documents is so great that it does not make sense to penalize for unaligned sentences. Instead, we divided by the number of sentences in the Easy Finnish document, because that would be the maximum possible number of alignments. We also did not take into account the probability that both documents were in the correct language because our task was monolingual.

It should be noted that we treated document and sentence alignments as exclusive. So, if document 1 aligned with document 2, no other document could align with documents 1 or 2. In all document alignment methods for documents from 2014 to 2018, we employed a simple strategy to find the best match for each document after obtaining the K best candidates through other methods. For all Easy Finnish documents, we found a maximum of five possible Standard Finnish matches, obtaining a matrix of distances or similarities. Then, we found the maximum (for similarities) or minimum (for distances, which is what the Vecalign method returns) value in the matrix. We locked that document pair, eliminated it from the matrix, and looked for the next highest or lowest value.

We used the manually-assessed document-aligned dataset with the articles from 2019 to 2020 to evaluate the quality of automatic document alignment for the years 2014-2020. During alignment evaluation, we only compared the document pairs

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

that were present in both the predicted sample and the annotated dataset, so the support was different in every case. When counting the "strict" scores, we considered "neutral" documents to be positive, and when counting "lax" (relaxed) scores, we considered the "neutral" documents to be negative.

We experimented with different thresholds for cosine similarity and distance scores. In our case, the distance was 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 were 9 possible thresholds from 0.1 to 0.9. We reached the conclusion that in the majority of cases, good F1 scores could 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 3.6 contains the evaluation results for the document alignment algorithm from Thompson and Koehn (2020) [the second approach], and Table 3.7 contains the results of document comparison with just cosine similarity between averaged sentence vectors [the first approach].

**Reading guide** for Tables 3.6 and 3.7:

- Emb (embeddings): which embedding model was used in the particular experiment. "-128" denotes truncated embeddings;
- Dist (Table 3.6): the cosine distance threshold between the document vectors (1 - cosine similarity). Pairs with cosine distance below this threshold are considered good matches;
- Cos. sim. (Table 3.7): the cosine similarity threshold between document vectors. Each document vector is an average of all sentence vectors. Pairs with cosine similarity above this threshold are considered good matches;
- Strict scores (precision, recall, F1-score): "positive" and "neutral" pairs in the reference dataset are considered "true";
- Lax scores (precision, recall, F1-score): only the "positive" pairs are "true";
- "Sup-1": support-1, the number of pairs deemed "positive" (true pairs) under the current threshold;
- "Sup-2": support-2, the number of document pairs in the predicted sample that match the document pairs in the reference dataset.

**Table 3.6.** [Table 1 from Paper VIII] Document alignment with Vecalign document embeddings (Thompson and Koehn, 2020).

| | | **Strict** | | | **Lax** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Emb** | **Dist** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0.9 | 0.723 | **1** | 0.84 | 0.82 | **1** | **0.901** | 1439 | 1439 |

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Emb** | **Dist** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| MPNet-128 | 0.9 | 0.718 | **1** | 0.836 | 0.814 | **1** | 0.898 | 1453 | 1453 |
| DistilUSE-128 | 0.9 | 0.712 | **1** | 0.832 | 0.808 | **1** | 0.894 | 1473 | 1473 |
| LASER-128 | 0.9 | **0.73** | 0.993 | **0.841** | **0.823** | 0.99 | 0.9 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0.9 | 0.728 | **1** | 0.842 | 0.824 | **1** | 0.903 | 1424 | 1424 |
| MPNet | 0.9 | 0.717 | **1** | 0.835 | 0.814 | **1** | 0.897 | 1473 | 1473 |
| DistilUSE | 0.9 | 0.711 | **1** | 0.831 | 0.807 | **1** | 0.893 | 1504 | 1504 |
| LASER | 0.9 | 0.729 | **1** | **0.843** | **0.826** | **1** | **0.905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0.701 | **1** | 0.824 | **0.80** | **1** | **0.89** | 743 | 743 |
| LASER rescored | n/a | **0.706** | **1** | **0.828** | 0.803 | **1** | 0.891 | 595 | 595 |

**Table 3.7.** [Table 2 from Paper VIII] Document alignment by comparing averaged sentence embeddings.

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Emb** | **Dist** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| LaBSE-128 | 0.68 | 0.717 | **1** | 0.835 | **0.812** | **1** | **0.896** | 1613 | 1613 |
| MPNet-128 | 0.55 | 0.701 | **1** | 0.825 | 0.797 | **1** | 0.887 | 1628 | 1628 |
| DistilUSE-128 | 0.47 | 0.689 | **1** | 0.816 | 0.783 | **1** | 0.878 | 1710 | 1710 |
| LASER-128 | 0.8 | **0.719** | **1** | **0.836** | 0.81 | **1** | 0.895 | 1574 | 1575 |

LaBSE and LASER embeddings gave 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 precision decrease, but we also got comparatively low support scores, which means that the set of document pairs that this algorithm retrieved matched the document pairs in the "true" data set rather vaguely. It can be seen that the candidate generation algorithm from Vecalign alone 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 truncated embeddings can be used in a more data-dense setting.

After evaluating different document alignment methods on the reference dataset (Dmitrieva et al., 2022), we used the Vecalign candidate generation method with full-size LASER embeddings to create the parallel document-aligned Finnish-Easy Finnish dataset for years, 2014-2018 (Dmitrieva and Yleisradio, 2024a).

#### 3.3.2.2 Sentences

After obtaining document-aligned data from the years, 2014-2020, the next logical step was to make sentence-level alignments. This format is more common for training text simplification models. However, in the case of our dataset, sentence alignment proved to be a challenging task.

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 the closest.

The first method that we used was **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 embedding model. It does not provide crossing alignments but satisfies all other requirements.

Our second aligner was **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 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 that we assembled, this was the maximum number of consecutive sentences appearing in one alignment, and 3:n and n:3 alignments were seen very rarely, so we did not see a reason to go over that limit. So, for example, between document 1 with sentences A, B, C and document 2 with sentences a, b, c, d, we could only align A to a, ab, or abc, but not to abcd.

We employed two baselines. The first was **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 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 used it with default values as in the example script[25], since we had discovered empirically that it is possible to obtain sensible alignments with these values. As a stop-words list, we used the stop-words list for Finnish from NLTK.

The other baseline that we used was a simple algorithm similar to the one described earlier in Subsection 3.3.2 for choosing the best documents out of K best. We embedded all sentences and concatenations of consecutive sentences (of length 1 <= len <= 3) and obtained a **cosine similarity matrix**. Then, we looked for the greatest value in this matrix, locked that alignment, eliminated all the sentences that went into that alignment (for instance, if we aligned sentences AB from

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

document 1 to sentence b from document 2, we also eliminated rows A, B, ABC, BC, ab, abc, bc, bcd), and looked for the next highest value. This method satisfied all our criteria.

For evaluation, we used the script provided in the Vecalign GitHub repository[26] 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 2019-2020 (Dmitrieva et al., 2022). There were 1638 singular sentences in Standard Finnish documents and 291 sentences in Easy Finnish documents. Between these documents, there were 223 non-zero alignments in the golden test set, of which 160 were one-to-one, 47 were one-to-many or many-to-one, and 16 were many-to-many ("many" was never higher than 3). The results can be seen in Table 3.8:

**Table 3.8.** [Table 3 from Paper VIII] Sentence alignment by different methods. Please refer to the reading guide in Subection 3.3.2 for column name explanations.

| | **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** | | | | | | |

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

| | Strict | | | Lax | | |
|---|---|---|---|---|---|---|
| **Embeddings** | **p** | **r** | **f1** | **p** | **r** | **f1** |
| n\a | 0.57 | 0.238 | 0.335 | 0.774 | 0.318 | 0.451 |

It can be seen that Vecalign with LASER embeddings outperforms all other methods. Bertalign seems to work significantly 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), which 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 large part of our data may be difficult to align even for humans. The greater the length difference between the Easy Finnish and Standard Finnish documents was, the harder it was to find true matches between the sentences.

Vecalign provides a score for all non-zero alignments, which reflects the cost of the alignment. The cost is calculated by the scoring function, which takes into account the cosine distance between two embeddings of blocks of one or more sentences from the source and target documents, scales it by the number of sentences in the blocks, and normalizes it based on the cosine distance between the blocks that are being considered for alignment at the moment and randomly sampled sentences from the entire documents. The smaller the number is, the better the alignment. Zero scores are given to zero alignments, i.e., 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.

### 3.3.3 Dataset statistics

The statistics of the Parallel Corpora of Finnish and Easy-to-read Finnish can be seen in Table 3.9. We only considered “positive” document and sentence pairs with the Vecalign score <= 0.65. If the score limit is lifted, the total number of non-zero pairs in the entire dataset would be 56088.

**Table 3.9.** [Table 4 from Paper VIII, some notations modified] Dataset statistics.

| | **2019-20** | **2014-18** | **Total** |
|---|---|---|---|
| **Documents** | | | |
| Pairs | 1257 | 7004 | 8261 |
| Words (Standard Finnish) | 471565 | 1700469 | 2172034 |
| Words (Easy-to-read Finnish) | 69179 | 402274 | 471453 |
| **Sentences** | | | |
| Pairs | 2994 | 8950 | 11944 |
| Words (Standard Finnish) | 41056 | 116684 | 157740 |
| Words (Easy-to-read Finnish) | 26699 | 80926 | 107625 |

An example of a dataset entry can be seen in Table 3.10. This excerpt is taken from the Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive 2019-2020.

**Table 3.10.** [Table 1 from Paper VII] An example of a dataset entry (Dmitrieva et al., 2022). 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).

| **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 lauantain | Öljyn hinta nousi enemmän kuin Iranin vallankumouksen tai Kuwaitin sodan alettua. Öljyn hinta lähti odotetusti jyrkkään | 0.84402 | Positive | Small difference in selkonews: last sentence |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | a tehdyt iskut... | nousuun markkinoiden avauduttua maanantaiaamuna... | | | |

## 3.4 How Simple are the Simple Parts in the Parallel Datasets?

In Chapter 2, Section 2.5 I have already mentioned that the linguistic properties of Simple Language described in the Easy/Plain Language guides are actually present in simple and simplified texts across domains and languages. However, it is not always clear to what degree a text has been simplified.

With RuAdapt being comprised of texts adapted for learners of Russian as a second language, it can be presumed that the authors adhered to the established criteria such as lexical minima during adaptation. Some of the chosen lexical simplification strategies (discussed in Paper IV and earlier in this chapter) also point at the adaptations being aimed primarily at the international audience interested in learning Russian. Whether or not these texts are suitable for other audiences that can benefit from Easy Language (such as children, people of old age, people with learning disabilities) is an open question, and the answer will probably vary for each individual text. Nevertheless, one advantage of the RuAdapt dataset is that the texts are marked by CEFR grade levels, however broad these markings may be. This gives the prospective reader a perception of their chance to succeed in understanding the text, and is also useful for automatic text simplification. For example, it can be beneficial if one wants to train a model capable of simplifying texts for different CEFR grade levels.

The audience of the Yle Uutiset selkosuomeksi is much broader. According to the interviews conducted by Kulkki-Nieminen (2010), the three main 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). With a target audience as broad as that, it is difficult to produce content that is equally easy for all audience members to understand. The Easy Finnish news is undoubtedly easier to understand than the "regular" Yle broadcasts, but this "easiness" is hard to measure. In Paper VII we

established that most of the Easy Finnish Yle articles adhered to the following Easy Finnish criteria from Selkomittari 2.0:

- The text contains mainly general vocabulary evaluated as familiar to thereaders.
- The text does not contain many long words.
- The text contains no figures of speech that require creative reasoning to understand (to chip away at something, brain drain, etc.).
- 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 that 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 Easy Finnish articles were not particularly easy to understand. For example, we have encountered complicated vocabulary (such as the word *amurinleopardikissapariskunta* [the pair of Amur leopards]), 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 merely summarized the original ones with no sentence-level simplification. Obviously, not everything needs to be simplified, and sometimes the authors of Easy Finnish texts will use complex words or constructions if they consider it necessary based on their expertise. Nevertheless, it is evident that the "easiness" level might vary from one article to another. In the absence of automated tools for measuring text complexity, it is hard to say where the Easy Finnish Yle articles fall on any kind of complexity scale, such as the aforementioned CEFR grade levels.

In conclusion, the corpora described in this chapter can be used for "general" text simplification, where the task is just to make the text more readable to the general audience. RuAdapt in particular can also be used to simplify texts for learners of Russian as a second language. However, establishing where exactly the "simple" parts of these corpora lie on the Easy to Standard Language spectrum is a question for future research.

# 4 Automatic Text Simplification

## 4.1 Model architectures

This chapter will cover different approaches to creating automatic text simplification systems and ways to achieve optimal performance in simplification, particularly for Russian and Finnish, so as to meet the second research objective and answer the final research question of this thesis (RQ3). I will detail the exploration of automatic text simplification on the datasets described in Chapter 3, as in Papers II and VIII. I will also describe the experiments with various methods for performing automatic text simplification with limited resources, such as transfer learning (Paper I) and controlled simplification (Paper V). Finally, this chapter will touch on the evaluation of simplified texts (RQ1) in Section 4.2, which outlines various methods for automatic evaluation of simplification systems' output.

Most of the work on this project was carried out from the beginning of 2020 to the end of 2023. During these years, deep neural networks were already widely used for most natural language processing applications. Also, large language models (LLMs) such as ChatGPT became available to the general public. Still, most of the methods described in this chapter utilize relatively small models by today's standards, because the size of state-of-the-art models is growing constantly.

As previously discussed, text simplification is most often viewed as a monolingual translation task. Therefore, the same approaches that are used for multilingual machine translation can also be employed for automatic text simplification. Over the course of this thesis project, we mostly experimented with sequence-to-sequence neural networks, i.e. those that have an encoder and a decoder part. These networks are designed to transform one sequence into another (hence the name) and are widely used for machine translation tasks. An example of such a model can be seen in Figure 4.1. This model is comprised of two recurrent neural networks (RNNs): the main feature of these networks is their ability to "memorize" the information from the entire sequence by utilizing a hidden state. In the figure, it can be seen that the encoder network memorizes a sequence in Russian, and that memory is passed on to the decoder network, which translates this information into English.

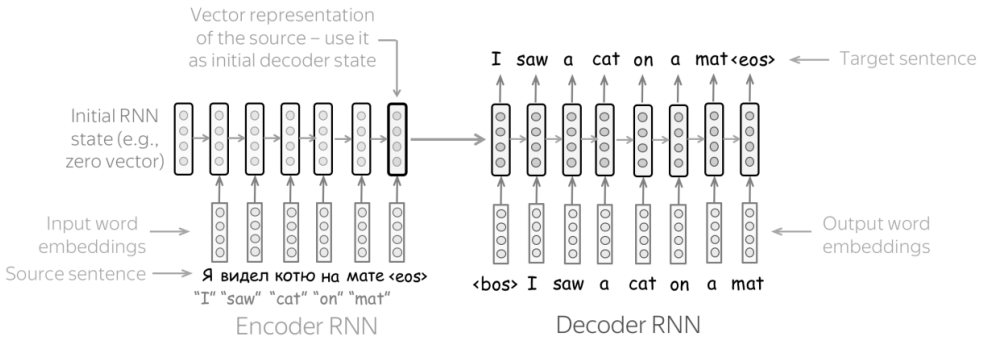


**Figure 4.1.** A sequence-to-sequence model with an encoder an a decoder part. (Voita, 2020)

In the first two papers, we mostly employed long-short term memory networks, or LSTMs (Hochreiter and Schmidhuber, 1997): a type of seq2seq-architecture that allows the model to "keep in mind" longer sequences while retaining only useful information and forgetting the rest. After that, we moved to transformer architectures, such as mBART and T5, which at the time were showing state-of-the-art results in tasks like summarization and simplification. In Paper VIII, we also employed Finnish GPT XL: a decoder-only model that was specifically developed for text generation tasks.

Transformer neural networks are also capable of "memorizing" an entire sequence, but in a more efficient and flexible way. Instead of the information being passed from the first token to the last, like in Figure 4.1, each token in a transformer "attends" to the current, previous, and next tokens and retains the information that is most relevant to the current token (this mechanism is called "attention" and is described in Vaswani et al. (2017). Transformers do not need recurrence to capture dependencies but can rely entirely on the attention mechanism.

For the experiments in Paper I, we used the Pointer-Generator model with coverage penalty proposed in Gehrmann et al. (2018)[27]. One of the main reasons for choosing this architecture is that, due to its data efficiency, this technique can be easily adjusted to a new domain (ibid.), which in our case is the simplification task. Since this architecture has already been used on the CNN/DailyMail dataset, we are using similar model parameters to those previously used for this dataset.

The model in Paper I was built using the OpenNMT toolkit (Klein et al., 2017). It uses a one-layer LSTM with 512 hidden states and an embedding size of 128. The encoder is a bidirectional LSTM with 512 hidden states (256 in both directions). The

[27] The "Models" section in the published version of Paper I contains an error that may lead readers to assume that we also used a content selection model in addition to the Pointer-Generator model, as had the authors of the cited work. It should be noted that we did not perform content selection as described in Gehrmann et al. (2018).

model uses copy attention (Vinyals et al., 2015) to copy words from the source. The copy loss is divided by the length of the sequence in tokens, which was proven by (Gehrmann et al., 2018) to generate longer sequences during inference. This model uses Adagrad optimizer, no dropout, and gradient clipping with a maximum norm of 2. At the inference stage, beam search with a beam size of 10 is used, because it has been found out that bottom-up attention requires a larger beam (Gehrmann et al., 2018). Multiple penalties are applied during inference: length penalty is used to encourage longer sequences, coverage penalty is used to avoid repetitions, and repeating trigrams are blocked.

In Paper II, we chose the architecture based on Nisioi et al. (2017), which has proven to perform well on the English simplification task. The implementation of this architecture is openly available online[28] and has some modifications that have further improved its performance (Cooper and Shardlow, 2020). Following this implementation, we used OpenNMT-py to build our models. As in the original paper, we used an architecture with 2 LSTM layers with hidden states of 500 and 500 hidden units. The dropout probability was set to 0.3. SGD was used as an optimizer, and global attention and input feeding were employed. We also employed the default learning rate of 1.0 with a decay of 0.7. The vocabulary size was set to 50000, which also happened to suit our needs, since the vocabularies of the original paragraphs exceed this number only slightly and the adapted vocabularies are even smaller.

As the project progressed, we moved to larger models. Unlike the neural networks described before, these models are usually not trained from scratch every time. Instead, on the first stage, these models are pre-trained on various tasks in order to gain language skills (often in multiple languages) and world knowledge. There are many such pre-trained models available online, many of them made by corporations such as Google, Microsoft, Meta, and the like. The users can download these models and fine-tune them to a particular task without repeating the expensive pre-training process, which requires a lot of data. In Paper V, we used versions of two transformer architectures, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020).

The first model that we used was mBART cc25 (Y. Liu et al., 2020), a model with 12 encoder and decoder layers trained on a monolingual corpus of 25 languages[29]. mBART was pretrained on the task of denoising full texts in multiple languages, which has allowed it to be directly fine-tuned for machine translation (ibid.). The preprocessing, training, and inference process in Paper V was identical to that of

[28] https://github.com/senisioi/NeuralTextSimplification

[29] https://github.com/facebookresearch/fairseq/blob/main/examples/mbart/README.md

the RuSimpleSentEval competition baseline[30]. We used the weights provided in the fairseq library[31].

The second model was a version of Google's multilingual T5 (Xue et al., 2021) with only Russian and some English embeddings left[32]. The original T5 model was pretrained on multiple transformation tasks, each converted into a text-to-text format (Raffel et al., 2020). The fine-tuning process of our model was similar to the one used by Дале (2021) for fine-tuning a T5 model for multiple tasks, including paraphrasing Russian texts (Дале, 2021).

The same mBART model was also used in Paper VIII. In addition to that, we also used the Finnish GPT-3: a Generative Pretrained Transformer with 1.5B parameters for Finnish (Luukkonen et al., 2023). We used the XL version[33] and fine-tuned it according to the authors' instructions[34]. It should be noted that GPT models are not sequence-to-sequence, but decoder-only architectures. Therefore, the task of translation is reformulated as a text generation task that follows a textual prompt. Nevertheless, with large enough model size and enough training data, these architectures have proven to be effective in simplification problems as well (Vadlamannati and Şahin, 2023).

One can employ different approaches to automatic text simplification. For example, if the available data is scarce, transfer learning could be used, in which a model would learn multiple tasks with an expectation that the knowledge would be transferred between the tasks. In Paper I, we explored the possibilities of combining the summarization and simplification tasks, expecting the model to use its summarization skills to aid the simplification process. In order to have more control over the model's output, various methods of controlled simplification can be utilized. For instance, in Paper V, I explored the possibilities of tailoring BART's output to match certain CEFR grade levels. Finally, larger models can learn instructions instead of just operating in a sequence-to-sequence setting. Instruction tuning can also give the user more control over the output. This method has not been fully explored during the work on this project, but we have touched upon it in Paper VIII.

Modern neural network models can produce high-quality simplifications, but even the largest models are prone to errors. The best way to assess the quality of a model is through human evaluation. However, such evaluation can be an expensive and lengthy process, sometimes worth a separate research paper. During the work on this thesis, we never evaluated the simplification models' output extensively with the help of human assessors, although we have always inspected the output

[30] https://github.com/dialogue-evaluation/RuSimpleSentEval
[31] https://github.com/facebookresearch/fairseq/blob/main/examples/mbart/README.md
[32] https://huggingface.co/cointegrated/rut5-base
[33] https://huggingface.co/TurkuNLP/gpt3-finnish-xl
[34] https://github.com/spyysalo/instruction-finetune

ourselves, albeit without performing any systematic qualitative assessment. Instead, due to time and resource constraints, we relied on language-agnostic automatic evaluation metrics designed for simplification evaluation, which are described in Section 4.2.

## 4.2 Metrics

At present, the most commonly used score for automatic simplification evaluation is **SARI**: a method for comparing system output against references and against the input sentence. It explicitly measures the goodness of words that are added, deleted, and kept by the systems (Xu et al., 2016). For each example, SARI takes into account the source text, the system's output, and at least one reference (target) sentence. It can also operate with multiple reference sentences. SARI takes into account the addition, deletion, and keeping operations happening between the texts. In the original implementation, the authors calculated the precision and recall for addition and keeping operations, and precision for deleting (because overdeleting hurts readability much more significantly than not deleting - ibid.). The score is calculated as follows (Xu et al., 2016):

$$SARI = \frac{1}{3}F_{\text{del}} + \frac{1}{3}F_{\text{del}} + \frac{1}{3}P_{\text{del}}$$

In general, SARI rewards operations that occur in output and in at least one of the references.

Before SARI became widely used, the **BLEU** score (Papineni et al., 2002) was often used for assessing simplification quality. BLEU is a method for automatic evaluation of machine translation that works by comparing system's outputs to a reference set of good quality human-made machine translations by calculating the proportion of n-gram matches between the outputs and references. However, BLEU has been recognized as being unfit for evaluating simplification (Elior et al., 2018), mostly because it does not account for sentence splitting properly, which is why we have only used it in some of the early papers in this thesis, mainly so that the scores could be compared to earlier works in simplification.

Another family of metrics that are sometimes used for simplification evaluation is **readability** scores. There are many formulas for automatic readability assessment, but perhaps the most popular score in the simplification community is the Flesch-Kincaid Grade Level, or **FKGL** (Kincaid et al., 1975). It is a simple score that takes into account the number of words per sentence and the number of syllables per word. The coefficients make the resulting score adhere to a certain level of education (a school grade or above) that is essential for understanding the text. It has been criticized for being too easy to manipulate (Tanprasert and Kauchak, 2021); however, it is still sometimes used in combination with other scores.

It should be noted that the exact scores that the system will get can differ depending on the framework that is used for evaluation. In all of the papers covered in this research, SARI and BLEU scores were calculated with the help of the EASSE library (Alva-Manchego et al., 2019). It can also be used for calculating FKGL; however, the constants in the FKGL formula need to be adjusted depending on the language, and the EASSE version is optimized for English, so we used a different implementation for Russian[35]. In some papers, additional scores were also employed.

It is worth mentioning that, in all of the experiments discussed below, we were interested in exploring a particular method rather than choosing the best hyperparameters for the models. Therefore, the evaluation focused on proving that the method was a viable option for implementing text simplification systems rather than finding the optimal configuration for the task. In particular, we wanted to test models that were trained from scratch, could be controlled in various ways, and systems that benefited from transfer learning and pre-trained models.

## 4.3 Methods and Results

### 4.3.1 Training a Neural Network From Scratch

Paper II provides an example of creating a simple automatic text simplification model from scratch. The primary goal of training that neural network was to test the newly created Russian-Simple Russian dataset to see if it was sufficient to create simplification models.

**Data:** the RuAdapt dataset is described in detail in the previous chapter (Section 3.2.1). At the time of working on Paper II, there existed only the adapted literature subcorpus of RuAdapt, which had two versions aligned with different aligners. The adapted literature subcorpus of RuAdapt that is currently available online is a little smaller than the one used in Paper II, since some novels could not be made public due to copyright reasons. Using Bleualign on our data resulted in 7452 aligned paragraphs, and with CATS-Align we obtained 9352. The test and development set sizes for the dataset aligned with Bleualign were both 1000 paragraphs, and for the data aligned with CATS they were 1500 paragraphs. Besides, a small test set of 302 manually aligned paragraphs from original and adapted texts for young native Russian speakers was used to test the model.

[35] https://github.com/infoculture/plainrussian

**Preprocessing:** the data was sufficiently cleaned before training. No further preprocessing techniques were applied, since our main goal was to see how good the quality of data was "as is."

**Experiments:** we used a simple LSTM network described in Section 4.1. This model was trained only on the RuAdapt data from scratch. Two models with the same architecture were trained separately on the same data and aligned with different aligners. The models were tested on the respective test sets, and also on the small manually aligned test set.

**Evaluation and results:** the results of our experiments are presented in Tables 4.1 and 4.2. At the time of writing this paper, there were no state-of-the-art Russian automatic text simplification systems yet, so we did not have anything to compare these scores to. However, the results we obtained demonstrate the success of the system even with today's standards, where a SARI score of around 40 is considered a reasonable outcome for Russian text simplification (see, for example, Sakhovskiy et al. (2021), and also the competition's GitHub: https://github.com/dialogue-evaluation/RuSimpleSentEval).

**Table 4.1.** [Table 5 from Paper II] Simplification evaluation – larger automatically aligned test sets.

| **Aligner** | **BLEU** | **SARI** | **FKGL** |
|---|---|---|---|
| Bleualign | 21.68 | 42.97 | 2.82 |
| CATS | 14.69 | 40.94 | 2.82 |

As can be seen in Table 4.1, models trained on the data aligned with Bleualign tend to have higher BLEU scores. However, the SARI scores for both datasets are close, and the outputs are equally readable according to FKGL.

As for the performance on completely unseen out-of-domain data, both models' performance declined. The best results for each system on the small test set are presented in Table 4.2. It can be seen that, although the BLEU scores halve compared to the bigger test set, the SARI and FKGL scores do not show such a rapid decline. In fact, the change in readability is very small in comparison to Table 4.1.

**Table 4.2.** [Table 6 from Paper II] Simplification evaluation – small manually aligned test set.

| **Aligner** | **BLEU** | **SARI** | **FKGL** |
|---|---|---|---|
| Bleualign | 10.86 | 35.53 | 3.33 |
| CATS | 7.51 | 33.84 | 2.72 |

Some examples of simplifications made by the best model can be found in Table 4.3. It can be seen that some simplifications, especially for the shorter sentences, happen to be identical to the target sentences. However, in some cases, the model simplifies the source text more than the original target sentence: for instance, in one of the examples in Table 4.3, the word надобно is substituted for a more frequently used synonym нужно, even though the substitution does not happen in the target sentence.

**Table 4.3.** Examples of simplification. English translations done by me with the help of Google Translate.

| | | |
|---|---|---|
| Source | — Чего тебе надобно, Алексеевна? — строго спросила Домна Замородновна. | — What do you need, Alexeevna? — Domna Zamorodnovna asked sternly. |
| Target | — Чего тебе надобно, Алексеевна? — строго спросила Домна Замородновна. | — What do you need, Alexeevna? — Domna Zamorodnovna asked sternly. |
| Output | — Чего тебе нужно, — строго спросила Домна Замородновна. | — What do you need, — Domna Zamorodnovna asked sternly. |
| Source | — И что же? — закричал я в нетерпении услышать конец. | — So what? — I yelled, impatient to hear the end [of the story]. |
| Target | — И что же? — закричал я. | — So what? — I yelled. |
| Output | — И что же? — закричал я. | — So what? — I yelled. |

### 4.3.2 Model Fine-tuning

In the later works, such as Paper V and VIII, we moved from smaller neural networks that can be trained from scratch to bigger networks that are usually fine-tuned on top of pre-trained models. Paper VIII in particular is an example of using models of different types to solve a sequence-to-sequence task. The main objective of these experiments was close to that of Subsection 4.3.1: we compared the performance of models without any specific optimization of the training procedures on a new dataset to see if the dataset was sufficient for training for the simplification task.

**Data:** we used the sentence-aligned version of the Parallel Corpus of Finnish and Easy-to-read Finnish for the experiments. The dataset is described in Section 3.3.

**Preprocessing:** the texts from the Yle archives contain very little noise as they were taken directly from the Yle database, which contains articles published on the

yle.fi website. Therefore, after aligning the sentences, no further cleaning was needed. No other specific preprocessing was needed for the planned experiments.

**Experiments:** the models used in this study were the mBART and Finnish GPT models described above in Section 4.1. We used the fairseq library (Ott et al., 2019) to fine-tune mBART, and the Huggingface Transformers library (Wolf et al., 2020) to fine-tune Finnish GPT. Because we used different architectures, the training approaches were also slightly different. For mBART, we utilized the standard schema for sequence-to-sequence training: supplying the source sentences to be used for encoding and the target sentences to be used for decoding. For Finnish GPT, we used an approach called instruction fine-tuning (Ouyang et al., 2022), in which a model is given an instruction, an input, and an expected output during training. The expectation is that the model will learn to generate an output in a specific way after being given an instruction, which is the same across all the training data, and some input. We used a simple instruction "Mukauta selkosuomeksi" [translate to Easy Finnish].

**Evaluation and results:** for evaluation, we again used the SARI score from the EASSE library. The evaluation results can be seen in Table 4.4. "Highest SARI" denotes the highest SARI score observed across all the training epochs; coincidentally, both models achieved their highest score on the 10th epoch, and the score declined afterwards.

**Table 4.4.** [Table 5 from Paper VIII] Model evaluation results for sentence simplification.

| | **Highest SARI** | **Epoch** |
|---|---:|---:|
| mBART | 37.612 | 10 |
| Finnish GPT | 44.63 | 10 |

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 Levenshtein 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 did not report the lexical complexity score because, to the best of our knowledge, it is not language-agnostic in the current EASSE implementation. For comparison, we provided the quality estimation values between the source and target documents. The values can be seen in Table 4.5.

**Table 4.5.** [Table 6 from Paper VIII] Quality estimation reports from EASSE. "Compression" stands for compression ratio, "Levenshtein" stands for Levenshtein similarity, and "Additions" and "Deletions" stand for additions and deletions ratio.

| Feature | mBART | FinnGPT | Target |
|---|---|---|---|
| Compression | 0.71 | 0.68 | 0.743 |
| Sentence splits | 0.828 | 0.831 | 0.875 |
| Levenshtein | 0.782 | 0.61 | 0.559 |
| Exact copies | 0.181 | 0.036 | 0.02 |
| Additions | 0.057 | 0.297 | 0.403 |
| Deletions | 0.339 | 0.559 | 0.618 |

On average, both mBART and Finnish GPT compressed the target sentences less than the actual test target sentences were compressed relative to the test source sentences. However, both systems came close to the correct amount of sentence splitting. The higher Levenshtein similarity, the number of exact copies, and the smaller number of additions and deletions lead us to believe that mBART is a more conservative model, which can explain its lower SARI scores.

Some examples of simplifications produced by the models can be found in Table 4.6. It can be seen that mBART 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 this 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 greatly. 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. However, the outputs are coherent and provide at least some degree of simplification.

The fine-tuned Finnish GPT can be found in the following repository: https://huggingface.co/Helsinki-NLP/simple-finnish-gpt3-xl. The fine-tuned 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.

**Table 4.6.** [Table 7 from Paper VIII] Example simplifications. Finnish texts are from news articles (copyright: Yleisradio), and English texts are translations of the sentences above.

| **Source** | **Target** | **mBART** | **Finnish GPT** |
|---|---|---|---|
| Text 1 | | | |
| Suomen Apteekkariliiton mukaan tilastot osoittavat, että mielialalääkkeiden euromääräinen myynti on kasvanut Suomessa noin yhdeksän prosenttia viime vuodesta. | Mielialalääkkeiden myynti on kasvanut Suomessa tänä vuonna. | Suomen Apteekkariliitto kertoo, että mielialalääkkeiden myynti on kasvanut Suomessa noin 9 prosenttia viime vuodesta. | Suomen Apteekkariliitto sanoo, että mielialalääkkeiden myynti on kasvanut apteekeissa noin 9 prosenttia viime vuodesta. |
| Translation | | | |
| 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. | Sales of mood medications have increased in Finland this year. | The Finnish Association of Pharmacists says that sales of mood medications have grown in Finland by about 9 percent since last year. | The Finnish Association of Pharmacists says that sales of mood medications have increased in pharmacies by about 9 percent since last year. |
| Text 2 | | | |
| Onnettomuus tapahtui vähän ennen puolta päivää paikallista aikaa Coraton ja Andrian välisellä rataosuudella Etelä-Italiassa. | Onnettomuus sattui lähellä Andrian kaupunkia. | Coraton ja Andrian välisellä rataosuudella Etelä-Italiassa on tapahtunut onnettomuus. | Onnettomuus tapahtui vähän ennen puolta päivää. |
| Translation | | | |
| The accident happened shortly before noon at the local time on the line between Corato | The accident happened near the city of Andria. | An accident has occurred on the track section between Corato and Andria in southern Italy. | The accident happened shortly before noon. |

| Source | Target | mBART | Finnish GPT |
|---|---|---|---|
| and Andria in southern Italy. | | | |
| Text 3 | | | |
| Sankarikoiria Suomessa on nimetty jo vuosia. | Suomessa on valittu sankarikoiria jo monta vuotta. | Sankarikoiria Suomessa on nimetty jo vuosia. | Sankarikoiria on Suomessa. |
| Translation | | | |
| Hero dogs in Finland have been named for years. | Hero dogs have been chosen in Finland for many years. | Hero dogs in Finland have been named for years. | There are heroic dogs in Finland. |

### 4.3.3 Multi-Task Learning

In Paper I we explored the possibilities of transfer learning for low-resource text simplification by jointly learning to summarize and simplify texts. At the time of writing this paper, we had already established that this thesis project would be focused on simplification for languages without large simplification datasets, and that we wanted to start with Russian. In 2020, there were no parallel datasets for Russian text simplification available online, whether with open or restricted access, so we made the decision to test our multi-task learning approach on English data, with the hopes of later experimenting with Russian texts.

Since simplification and summarization are very similar tasks, they are sometimes combined in the development of systems for text complexity reduction. (Saggion, 2017) lists multiple examples of such systems. In one of them, sentence simplification was used as a part of a multi-document extractive summarization algorithm. Sentences were simplified before clustering and selecting relevant sentences from each cluster (ibid.). In another system (Lal and Ruger, 2002), lexical simplification is used during summary generation to replace difficult words.

The decision to combine the summarization and simplification tasks was made in part because we believed that, in most languages, summarization datasets are easier to build than simplification datasets: for example, by aligning Wikipedia article summaries (short paragraphs that preface the main body of text in an article) with the text of the article. Simplification-specific datasets, on the other hand, are difficult to build, since a lot of languages such as Russian do not have "simple" versions of regular media sources. For example, while the "regular" Wikipedia exists in many languages, Simple Wikis are quite rare. Besides, both simplification and abstractive summarization require a good understanding of text semantics, so we

assumed that finding a proper way to preserve the meaning of texts could benefit both tasks.

As mentioned above, we used the OpenNMT-py toolkit (Klein et al., 2017) to preprocess our data, as well as to train and test the models.

**Data:** for the purposes of this research, we used two different datasets: the Simple Wikipedia dataset (Kauchak, 2013) for simplification and the CNN/DailyMail dataset (Nallapati et al., 2016) for summarization. We used data aligned at the document level instead of a sentence-to-sentence alignment approach, which at that time was (and still is) commonly used in simplification. First, document-level alignment is common for text summarization datasets as it allows models to learn to omit entire sentences and parts of the text. Second, for languages without large simplification datasets, it would be easier and less time-consuming to create a simplification dataset aligned at the document level. CNN/DailyMail has around 300 thousand documents (231 mln. tokens), and SimpleWiki has 60 thousand (94 mln. tokens).

**Preprocessing:** we employed multiple methods to preprocess the texts. First of all, punctuation was detached from words, as it was required for the type of models that we trained. Secondly, for some of the models, we applied an additional sentence tokenization method. It had been shown that the summarization models can perform better on the CNN/DailyMail dataset if sentence boundary tagging is applied to the target text beforehand like this: ‹t› w1 w2 w3 . ‹/t› (Gehrmann et al., 2018). We also used two types of word tokenization: singular words as tokens and subword units obtained from SentencePiece[36]. The latter allowed us to take advantage of a smaller vocabulary.

We followed the approach of See et al. (2017) for automatic summarization on CNN/DailyMail and truncated the source texts to 400 tokens and the target texts to 100 tokens. Although such truncation might seem brutal, See et al. (2017) proved that, at least for the CNN/DailyMail dataset, it can improve the performance of the summarization model. We also used a dynamic dictionary and shared vocabulary to ensure that source and target sentences were aligned and used the same dictionary, which is needed for copy attention – an implementation of pointer-generator networks that considers copying words from the source sequences (Gehrmann et al., 2018; Wang et al., 2016).

**Experiments:** we tested the following setups to determine the best approach:

- **Experiments on a singular dataset.** In order to evaluate the performance of the models trained on the two datasets correctly, the models were also trained on each dataset separately. Moreover, separate

[36] https://github.com/google/sentencepiece

models were trained on versions of the datasets with and without sentence boundary tagging. In addition, we also experimented with applying different SentencePiece tokenization and comparing the performance of models trained on data preprocessed with SentencePiece to that of models trained on regular data.

- **Experiments with fine-tuning.** We first trained a model on the summarization dataset and then fine-tuned it on the smaller simplification dataset. For this experiment, we also employed different versions of the dataset, with and without sentence boundary tagging.
- **Experiments on the joined dataset.** Some models were trained on both datasets simultaneously. All source texts were augmented with task-specific tags: "<2sum>" at the start of a sentence meant that the text was to be summarized, and "<2simp>" meant that the text was to be simplified. We tried multiple approaches to compensate for the mismatch in the sizes of the datasets. In one experiment, the model was trained on a joined dataset with the volume of each original dataset preserved. In another experiment, the summarization data was undersampled and the simplification data was oversampled (with some texts being repeated) so that the amount of source simplification data was 117,550 texts and the summarization data was 287,227 texts. In the third experiment, the model mentioned in the first experiment was additionally fine-tuned on simplification data that it had already seen.

**Evaluation and results:** we used the BLEU, SARI, and FKGL (Flesch-Kincaid Grade Level) scores for evaluating simplification and readability, and also a pure Python implementation of the ROUGE score[37], which is a commonly used score for summarization evaluation.

Experiments on single datasets showed that the best simplification scores were obtained with the simplest preprocessing strategies, without SentencePiece or sentence boundary tagging. However, using BPE tokenization seems to give better results in summarization for models trained on a single dataset. Sentence boundary tagging, although it improved summarization performance on the CNN/DailyMail dataset, did not seem to be effective on other data.

Training the model on one dataset and then fine-tuning it on another proved to be less effective than using the datasets jointly with task-specific tags. As for joined training, the oversampling and undersampling approaches proved to be less effective than just combining the datasets as is without further fine-tuning. The models trained on the joined dataset could perform different tasks with the same

[37] https://github.com/pltrdy/rouge

effectiveness as models trained for one task and tested on the corresponding data. The different tokenization approaches tested did not seem to have a significant effect on the performance of these models.

Finally, we compared the performance of the models trained on the joined dataset on regular test sets and on test sets with reversed tags (where <2sum> becomes <2simp> and vice versa) in order to see if the models understood the semantics of task-specific tags. Evaluation scores decreased somewhat but the difference was less than expected. The length of the output texts, however, increased slightly on average, which was reflected in increased readability scores. A look at a small number of randomly selected output texts confirmed that, on average, the simplified articles were longer than summaries, even though all target texts were truncated to the same length during preprocessing. Examining the output also confirmed that the same source text with different tags were processed differently by the model. This can be illustrated with the example below (taken from the Simple Wiki dataset):

> **Original text:** sofia wistam -lrb- born 15 may 1966 in liding, stockholm county, sweden -rrb- is a swedish television host on tv4 and tv3 and radio talkshow host. she has also worked as a stylist for stars such as carola, jerry williams and tommy nilsson. in, 2008 she was also a judge on the talent show sweden's got talent, during this year she also hosted her own radio show on rix fm. during, 2009 sofia will host the competition show on swedish television.
>
> **Target text:** sofia wistam -lrb- may 15 1966 -rrb- is a swedish television host and radio talk-show host.
>
> **Output with <2simp> tag:** sofia wistam -lrb- born 15 may 1966 in liding, stockholm county, sweden -rrb- is a swedish television host on tv4 and tv3 and radio talk-show host. she has also worked as a stylist for stars such as carola, jerry williams and tommy nilsson.
>
> **Output with <2sum> tag:** sofia wistam is a swedish television host on tv4 and tv3 and radio talk-show host. she has also worked as a stylist for stars such as carola, jerry williams and tommy nilsson.

The outputs with different tags are not the same even when the source text is quite short. However, it is hard to pinpoint the exact differences that each tag triggers, not only because the evaluation of such phenomena is generally difficult, but also because the tasks of summarization and simplification are less distinguishable in nature than, for example, the tasks of translating a text into two different languages.

After creating a dataset specifically for Russian simplification, we discovered that the multi-task approach might not be needed for performance improvement.

However, we used some of the techniques, such as data augmentation, with datasets for similar tasks and employing control tokens in Paper V.

### 4.3.4 Controlled Simplification

The task-specific tags method described in the previous section can also be viewed as a method of controlling the output. In Paper V, we investigated the possibilities of controlling the output of simplification models further, this time focusing on tailoring the output to different linguistic properties.

There are multiple ways to control the output of text simplification tools. For example, editing operations can be controlled directly. Dong et al. (2019) presented a simplification model that could learn explicit editing operations such as additions, deletions, and keeping. Alva-Manchego et al. (2017) proposed a sequence labeling model to predict which simplification operations should be performed as a first step for a complete simplification pipeline. The model was built on a corpus with automatically labeled simplification operations, and the approach has proven to produce more straightforward texts than end-to-end models. More recently, Cripwell et al. (2023) have described a way to perform controlled simplification on the document level by first generating a document-level plan that stipulates a sequence of sentence-level simplification operations (copy, rephrase, split, or delete) for the input document, then using this plan to guide the iterative generation of the simplified document across sentences.

Other research shows that, apart from controlling editing operations, it is also possible to control specific dimensions of the output texts. Martin et al. (2020) identify four attributes related to the text simplification process: the amount of compression, paraphrasing, lexical and syntactic complexity – and use control tokens that are put in front of the source sentences to modify these attributes in output texts. This approach was later used in Martin et al. (2022) and in Anastasyev (2021). The latter was the winning solution for the RuSimpleSentEval (Sakhovskiy et al., 2021) shared task on Russian text simplification. This methodology was used in Paper V as well. Other studies have shown that control tokens can be used for all kinds of linguistic attributes, including politeness and monotonicity (the closeness of the word order in the target sentence to the word order in the source sentence) (Schioppa et al., 2021), and even psycholinguistic features such as prevalence, which refers to the number of people who know the word (Qiao et al., 2022). Some studies also demonstrate the successful use of control tokens to generate texts for a given school grade level (Scarton and Specia, 2018; Nishihara et al., 2019).

**Data:** four different data sources were used in Paper V. I developed on the ideas from Paper I and included some datasets for tasks that are similar to simplification, such as paraphrasing, in order to augment the simplification data:

- ParaPhraser Plus: a large automatically developed corpus for Russian paraphrase generation (Gudkov et al., 2020). Contains news headlines crawled from publicly available websites;
- Opusparcus: a paraphrase corpus for six European languages comprising subtitles from movies and TV shows (Creutz, 2018). Only the Russian part of the corpus was used;
- RuAdapt: a parallel Russian-Simple Russian dataset which consists of texts adapted for learners of Russian as a foreign language (see Paper II and the previous chapter). As mentioned before, sentence pairs in RuAdapt were aligned automatically and have cosine similarity scores provided by the aligner. Only sentences with cosine similarity above 0.31 but below 0.98 were used;
- The RuSimpleSentEval[38] datasets: development and public test set (Sakhovskiy et al., 2021). The original training set was unavailable at the time of writing the paper. The public test set of 3398 sentence pairs was only used separately from the rest of the data.

All data used was aligned on the sentence level, although in some cases there could be one-to-many, many-to-one, and many-to-many alignments. In total, there were 455,327 sentence pairs in the training set, 50,592 in the development set, and 10,325 pairs in the test set, not including the RuSimpleSentEval public test. The resulting dataset was then preprocessed in a way that would ensure that the source sentences were more complicated than the target ones.

**Preprocessing:** only sentences with five tokens or more were used in this study, because accurately estimating CEFR grade level for very short sentences is impossible. Furthermore, to avoid incoherent, ungrammatical outputs and hallucinations (content that is inconsistent with the real world or user input – X. Liu, 2024), the larger parts of the dataset, Paraphraser Plus and Opusparcus, were cleaned of sentence pairs in which named entities did not match. The Natasha toolkit[39] was used to exclude sentence pairs where the target sentence had named entities that were absent in the source.

The source texts were augmented with the following control tokens as per Martin et al. (2022) and Martin et al. (2020):

- **NbChars**: the ratio between the lengths of source and target sentences in characters; represents the amount of compression. Same as in Martin et al. (2020);
- **LevSim**: the Levenshtein ratio between source and target sentences; represents the amount of paraphrasing. Same as in Martin et al. (2020);

[38] https://github.com/dialogue-evaluation/RuSimpleSentEval
[39] https://github.com/natasha/natasha

- **DepTreeDepth**: the ratio between the syntactic tree depths of target and source sentences; represents the syntactic complexity. Similar to Martin et al. (2020). The dependency parsing was performed with the deeppavlov[40] ru_syntagrus_joint_parsing model;
- **CEFRgrade**: the CEFR grade level of the target sentence; represents multiple simplification-related attributes. It is the only token not represented by a ratio because it is easier to control the output's grade level directly rather than control how simplified the output will be compared to the source. The grade levels were calculated using code from the Textometr's (Laposhina et al., 2018) API. Textometr's grade levels go from elementary A1 up to what can be described as C2+ (too complicated even for a native speaker) and can be transformed to a 0.0 to 10.0 scale. Only sentence pairs in which the source's grade level was higher than or equal to the target's (which means that some pairs had to be reversed) and the target's CEFR level was not higher than C2 were kept in the dataset.

Here is what a source sentence with control tokens could look like before encoding and preprocessing (this sentence is from the ParaPhraser.ru corpus):

<CEFRgrade_0><LevSim_0.4><NbChars_1.15> Погода на завтра: преимущественно без осадков.

*Weather for tomorrow: mostly without precipitation.*

Previous research had shown that the NbChars and LevSim tokens worked well for both English and Russian; therefore, they were chosen for the initial experiments, including experiments with choosing the model architecture. To the best of our knowledge, the DepTreeDepth token had never been tried on Russian but had shown a slight performance increase for English (Martin et al., 2020), so it was included in later experiments. The reasons for choosing CEFR grade level as one of the tokens were twofold. The first goal was to find a way to simplify texts for a particular grade level. Secondly, since the WordRank token used in Martin et al. (2020) did not work well for Russian (Anastasyev, 2021), it was necessary to find something else to represent the change in lexical (and other) complexity between sentences. Moreover, studies such as Scarton and Specia (2018) had shown that annotating the source sentences with information about the target grade level can positively affect the model's simplification performance. All tokens except CEFRgrade levels had 40 unique values from 0.05 to 2. The tokens were appended to the beginning of the sentence. Then the sentence was encoded with SentencePiece, preprocessed with fairseq, and fed to the model. Therefore, no

[40] https://github.com/deeppavlov/DeepPavlov

special embeddings just for the control tokens were added to the pretrained model, and the vectorization of control tokens happened as is. This approach was used in Martin et al. (2022). It would have been interesting to experiment with appending new embeddings to the pretrained models to represent control tokens. For instance, in Schioppa et al. (2021), the authors introduced attribute control during fine-tuning by affecting a smaller subset of the original model parameters. However, at the time of writing Paper V, most frameworks did not have such functionality.

**Experiments:** as mentioned above, we used mBART and T5 for experiments with controlled simplification. The models were evaluated on two test sets: the general test set and the public test set from RuSimpleSentEval (RSSE).

The models' performance was evaluated with the SARI score from the EASSE library. Before evaluation, sanity tests were conducted on the RSSE public test set: if the source file was used as the output file, the SARI score was 14.7, and if the target is used was output, the score was 100. During RuSimpleSentEval, the best system had a SARI score of 40.23 on the public test set.

**The following experiments were conducted:**

- **Training without any control tokens.** This was the starting point of the experiments, which was necessary for determining if the control tokens actually aid in the training process.
- **Training with the NbChars and LevSim tokens.** These tokens had already shown good performance on English and Russian data, so they were a good starting point to check whether the experiment has been set up correctly and whether these results could be replicated on our data. At this stage, we performed multiple experiments with the values of the control tokens. First, we set the values to 1.0: the hyphothesis was that if the models "understood" the meaning of the control token values, the source sentences would be left unchanged. To further investigate how the control tokens affected the model, we measured the actual values of the character length ratio and the Levenshtein similarity ratio between the model's output and the source sentences. Intuitively, suppose a model was asked to simplify sentences with NbChars set to 0.95. In that case, the average character length ratio between the system output and source sentences should be close to 0.95. These experiments also helped us to choose between the two model architectures that we initially employed.
- **Training separate models with only "new" tags** (never before used on Russian data): DepTreeDepth and CEFRgrade. Experiments with the latter token also involved determining the optimal number of possible values, since CEFR grade levels had not been used as control tokens prior to Paper V, and testing the influence of different control token values on

the output: for example, whether or not the outputs really belonged to the CEFR grade level that was set by the user during inference.

- **Combining** multiple control tokens in one model.

**Evaluation and results:** when trained without any control tokens, mBART had a much higher score on the general test set, but on the RSSE (RuSimpleSentEval) public test set, the scores were much lower, with T5 performing slightly better. However, adding two control tokens, NbChars and LevSim, improved the performance of mBART significantly on both test sets. T5, however, did not show a considerable performance gain. Moreover, when both tokens were set to 1.0, only mBART showed a SARI score similar to the SARI that can be obtained if the source sentences are passed as output. It should be noted, however, that, despite high SARI scores, the output of mBART contained some wrong (in relation to the source), incoherent, and/or ungrammatical sentences. Anastasyev (2021) also reports that the models with highly rated performance still hallucinated in some cases, although the type of hallucination was not specified.

When we compared actual values of the character length ratio and the Levenshtein similarity ratio between the model's output and the source sentences, we found that models seemed to learn the meaning of the tokens with further training, even though it did not necessarily mean SARI score improvement. Evidently, the mBART architecture was better at understanding the meaning of both control tokens, which is why it was chosen for further experiments. It should also be noted that the training process for mBART with fairseq was faster than training T5 with transformers, which influenced our choice of model.

Training an mBART model with the same configuration as before on texts with just the DepTreeDepth token resulted in a considerable decrease in performance. After 5 initial epochs and an additional 7 epochs after early stopping, the best SARI score on the general test set was 28.77 on epoch 7. Despite generally standard loss scores (not much different from previous experiments with and without control tokens), the models hallucinated frequently. The hallucinations made calculating the actual syntactic tree depth of the outputs impossible because there were too many word repetitions to create adequate syntactic trees. In conclusion, the tree depth ratio may not be an adequate metric to control syntactic complexity in Russian sentences. It should be noted that, as reported in Martin et al. (2020), the identical DepTreeDepth token also did not seem to control its attribute as well as the NbChars and LevSim tokens did in English texts, although it had the desired effect on the output.

In order to train mBART to understand CEFR grade levels, we first conducted multiple experiments to determine how many unique values should be allocated to this token. The starting range was from 0.7 to 8.5 with a step of 0.1 (how the values come from Textometr). After a decrease in performance compared to models with

no tokens (the highest SARI score obtained on the general test set was 35.84 on epoch 8/12), the number of unique values was lowered to 8, from 1 to 8. After that, the SARI scores increased up to around 41 (epoch 4/7), but the model still hallucinated quite a lot. Consequently, the number of unique values was reduced to 6, corresponding to levels A1 (0) to C2 (5). This decreased the SARI scores slightly (highest SARI 38.97, epoch 8/10); however, the outputs became more coherent.

In order to test the influence of different token values on the output, during the inference, the token was set to lower grade levels, from A1 (0) to B2 (3). The testing has shown that the SARI score decreases when the CEFR grade level goes up. As expected, the lowest CEFR grade gave the highest SARI score (46.5 with CEFR grade set to 0, 38 with CEFRgrade set to 3, 39 when set to the target's actual CEFR grade level). When studying this token's influence further, it became clear that, even though setting the token to a particular grade level leads to more sentences of that level in the output, the model still produces a lot of B1 and B2 (2 and 3) level sentences. The reason is likely because there are many sentences with these grade levels in the training data.

Despite the model being able to learn the NbChars and LevSim control tokens together and the CEFRgrade separately, combining them in one model did not increase performance. On the contrary, there was no noticeable SARI increase across 18 epochs, and many outputs were incoherent, with a lot of word repetitions. The reason for such behavior is unclear, since in previous studies (see, for example, Martin et al. (2022), and Schioppa et al. (2021), different control tokens were successfully combined.

In conclusion, the experiments have shown that the DepTreeDepth token does not perform as well on Russian data as it did on English, according to previous research. Therefore, some tokens are "harder" for the models to learn than others. Nevertheless, the CEFRgrade token can influence the model's output in a desirable way, but according to the results of the experiments, combining it with other tokens worsens the model's performance. Moreover, the probability of getting outputs with the desired CEFR grade level will be influenced by the dataset composition: the model will gravitate towards producing sentences on the level that it saw the most in the target sentences. Finally, we have confirmed that the other two tokens, NbChars and LevSim, work well on Russian data. Some examples of simplifications performed by the models can be found in Tables 4.7 and 4.8. The best models' checkpoints and other supplementary materials can be found on GitHub: https://github.com/annadmitrieva/controlled_simplification_ru.

**Table 4.7.** [Table 6 from Paper V] Examples of simplifications with arbitrary CEFR grade levels. Original dataset: ParaPhraser.ru. English translations done by me with the help of Google Translate.

| **Partition** | **Text/translation** | **Actual grade level** |
|---|---|---|
| Source | Семья Березовского не дает согласия на закрытие уголовных дел против него | 3 |
| | Berezovsky's family does not consent to the closure of criminal cases against him | |
| Target | Родственники Березовского не будут давать согласие на прекращение уголовных дел в отношении него | 3 |
| | Berezovsky's relatives will not consent to the termination of criminal cases against [in relation to] him | |
| $CEFRgrade_0$ | Семья Березовского не хочет закрывать дела | 0 |
| | Berezovsky's family does not want to close the cases | |
| $CEFRgrade_1$ | Семья Березовского не хочет закрывать дела против него | 1 |
| | Berezovsky's family does not want to close cases against him | |
| $CEFRgrade_2$ | Семья Березовского не дает согласия на закрытие уголовных дел | 2 |
| | Berezovsky's family does not consent to the closure of criminal cases | |
| $CEFRgrade_3$ | Семья Березовского не согласна на закрытие уголовных дел против него | 3 |
| | Berezovsky's family does not agree to the closure of criminal cases against him | |

**Table 4.8.** [Table 7 from Paper V] Examples of simplifications with arbitrary NbChars and LevSim parameters. Original dataset: RuSimpleSentEval public test. English translations done by me with the help of Google Translate.

| **Partition** | **Text/translation** |
|---|---|
| Source | Андропов,военный атташе и водитель уцелели и пешком добрались до посольства. |
| | Andropov, the military attache and the driver survived and reached the embassy on foot. |

| Partition | Text/translation |
|---|---|
| Target | Андропов вместе с военным атташе и водителем уцелели, но пешком два часа по ночному городу пробирались в посольство. |
| | Andropov, along with the military attache and the driver, survived, but they made their way to the embassy on foot for two hours through the night city. |
| $NbChars_{1.0}$, $LevSim_{1.0}$ | Андропов,военный атташе и водитель уцелели и пешком добрались до посольства. |
| | Andropov, the military attache and the driver survived and reached the embassy on foot. |
| $NbChars_{0.95}$, $LevSim_{0.4}$ | До посольства добрались Андропов, атташе и водитель. |
| | Andropov, the attache and the driver reached the embassy. |

## 4.4 Discussion

In this chapter, I described some methods and models for text simplification in low-resource settings. As the work on this project progressed, the models gradually became larger. At present, researchers in all fields of NLP are actively exploring the use of LLMs to solve various tasks, including simplification with fine-tuning or prompt engineering. Models such as mBART or Finnish GPT XL can now be considered "smaller", and today's "bigger" models can have tens of billions of parameters. For instance, Finnish GPT XL has 1.5B parameters, which translates into about 6 GB of disk space, and Finnish GPT 13B would occupy about 55 GB. While training these models requires a lot of resources (for example, it cannot be done on most personal computers), most of the time the model performance improves significantly when the number of parameters increases. "Smaller" models are still used for simpler tasks, or in cases when computational resources are scarce or a more environmentally friendly solution is needed.

The size of the model and the availability of computational resources also affects the context size that the model can process effectively. During this project, we mostly operated on the sentence level when fine-tuning, and on the paragraph or text level when training small models from scratch. In our case, this was enough for tasks such as testing models' performance on a new dataset or testing a new modeling technique. However, a "real world" end-to-end text simplification application would most likely operate on the document level since simplification requires a lot of contextual knowledge for summarizing or rearranging fragments of texts.

The problem of data scarcity for text simplification still persists today: even though new datasets are constantly appearing, we are still very far from being able to make a good simplification model for every language (even if endangered languages are not taken into account). At present, techniques like data augmentation and multi-task learning are still being utilized in low-resource settings. However, such techniques can present a risk of contaminating the data with pairs where the "simple" text is actually not simpler or even less simple than the original text. In Paper V we mitigated this problem by using Textometr to automatically determine the CEFR grade level of both texts in a pair. We then reversed the pairs where the target text had a higher grade level than the source, and then eliminated the pairs where the target text's grade level was above C2. This solution, however, is language-specific and also task-specific, since Textometr and the CEFR grade level system itself were made primarily for use in second language teaching.

Most models that are trained or fine-tuned for simplification will probably be audience-specific, since simplification strategies for different audiences can vary greatly. This is why being able to control the output of a simplification model is a desirable feature. Our experiments have shown that even "smaller" models can be trained to distinguish between certain degrees of simplification and perform the task accordingly. However, testing this technique on texts from different domains would require data that is not currently available for a lot of languages. For larger models, the method that we used would most likely be substituted by instruction fine-tuning or just prompt engineering.

# 5 Conclusions and Future Work

In this thesis project, I explored the task of automatic text simplification, focusing on the Russian and Finnish languages. In both cases, the work on simplification tools for the language began with making suitable datasets from scratch and concluded with developing multiple simplification models, thereby achieving the two main research objectives of this study. In addition, I studied the linguistic characteristics of simple language, as well as the relationship between the fields of language technology and accessible communication.

The practical contributions of this work include:

1. Creating multiple parallel datasets suitable for training text simplification models and studying the linguistic properties of simplified texts (Paper II, Paper IV, Paper VII, Paper VIII);
2. Exploring ways of performing automatic text simplification in low-resource settings (Paper I);
3. Exploring ways of controlling the output of simplification models (Paper V);
4. Training and fine-tuning simplification models of various kinds and performing model evaluation with automatized metrics (Paper I, Paper II, Paper V, Paper VIII).

The theoretical contributions of this work are as follows:

1. Defining the general principles of simple language and determining the degree of simplification in the "simple" parts of the parallel datasets created over the course of this project;
2. Studying the linguistic properties of simplified texts and simplification strategies (Paper III);
3. Describing the interaction between language technology and accessible communication (Paper VI).

This work focused on simplified texts and the process of simplification rather than Easy and Plain Language as a broader topic. In this study, a text was considered simplified or a model was considered able to simplify if the simplification could be measured in any way. Various automatically calculated metrics and visual examination of select examples served as proxies for determining that a text had become easier to understand. Although this approach is common in the academic community, it is obviously not ideal for measuring something as complex as simplicity for multiple reasons.

The **first research question (RQ1)** of this thesis was about defining and evaluating simplicity and simplification. Since this thesis is primarily concerned with automatic text simplification, this question was of particular interest in the context of developing ATS systems. In order to tackle RQ1, I compared seven Easy/Plain language guides for different European languages (and some language-agnostic) and outlined the general principles of simple language. I also described some of the strategies used by authors of simple texts and presented quantitative analysis of the simplified texts' linguistic features. Finally, I listed the automatic evaluation metrics suitable and most often used for evaluating simplification systems, and evaluated my own simplification models using them, along with presenting examples of the models' outputs.

First of all, easiness is not easily defined. Like most concepts in linguistics, easiness or simplicity in language does not have an unambiguous, universally accepted definition. It is also hard to describe empirically. As explained in Chapter 2, there are many criteria that define simple language, many of them language-specific. Also, the prospective audience plays an important role in establishing if a text is easy enough to understand. Therefore, a human simplicity evaluation would also most likely be incomplete and/or subjective. So, a "perfect" evaluation schema for text simplicity will likely never exist.

Secondly, while there can be no perfect evaluation tool, the evaluation can certainly be improved with the involvement of human experts or data-driven methods that can mimic a human's judgment of simplicity well. Unfortunately, human evaluation is expensive and time-consuming (especially considering the time needed to define and test the evaluation criteria), and data-driven tools are not truly language-agnostic. That is why, at present, most researchers use relatively simple language-agnostic metrics, especially when it is important to be able to compare results – for example, in model evaluation. In the future, however, it is possible that model evaluation might move to data-driven methods, at least for languages with enough resources.

Nevertheless, automatic metrics can be used to describe or evaluate at least some properties of simple texts. For example, in Paper V we used Textometr to estimate the CEFR grade levels of the training texts, and the model trained on this data actually learned to distinguish between several degrees of simplification. Textometr is not a fully data-driven tool in the sense that it is algorithm-based rather than machine learning-based, but it is language-specific. It makes use of many Russian word lists, such as frequency lists and lexical minima, many of which were also used to study the properties of texts simplified for different audiences in Paper III. The writing strategies that were identified automatically matched the intuitions that we had (such as texts for speakers of Russian as a second language having the most words from lexical minima or texts for native speakers having the most adjectives), and they also matched some of the criteria found in Simple Language guidelines.

This, and the successful modeling experiments, proves that simplicity can be identified and reproduced automatically. Obviously, some properties are easier to identify and reproduce than others, but modern language models are getting close to understanding even the more difficult concepts such as logic and consistency every day.

The **second research question** explored in this thesis **(RQ2)** dealt with the most effective strategies for creating text simplification datasets from various data sources and domains. To explore RQ2, we built parallel simplification datasets for Russian and Finnish languages from scratch, experimenting with various alignment strategies.

Chapter 3 contains detailed descriptions of how both the Russian and the Finnish parallel datasets were created. Because for both languages there are few (for Russian) or no (for Finnish) other parallel simple language - "regular" language datasets, it is not possible to make a thorough comparison between our datasets and other data sources. However, the experiments described in Chapter 4 have proven that the datasets are sufficient for training and/or fine-tuning simplification models. It should be noted that we have always used data created by professional authors: journalists, classical writers, or specialists in teaching Russian as a second language, so all the texts that our datasets are based on are of high quality. This factor can significantly influence the quality of the resulting datasets and simplification models.

For both datasets, we tested at least two different alignment methods. Since the automatic alignment techniques evolve constantly, the methods that we applied for the Finnish-Easy Finnish datasets were different from the libraries that we used for RuAdapt. Nevertheless, while working on the Finnish-Easy Finnish dataset, we came to the conclusion that the best solution for automatic alignment should be determined on case-by-case basis: for instance, while comparing the results that we obtained in Paper VIII to the results that other researchers had achieved with the same aligners on different languages, we found out that their best aligners were not the best in our case. It would be ideal if we could also publish the entire original and simplified texts that we used to create our datasets: it would be a good source for document-level simplification, and would give other researchers an opportunity to re-align the texts with other instruments if they want. Unfortunately, publishing a document-aligned version was only possible for the Finnish-Easy Finnish dataset.

The **last research question (RQ3)** was concerned with whether simplification models can be successfully trained on limited data in morphologically rich languages. To answer it, several simplification models were built on various corpora (including those described in Chapter 3). We have experimented with different model architectures and also different methods for performing automatic text simplification with limited data.

Because at the time of working on this thesis project there existed very few or no specialized simplification models for our chosen languages, it was hard to compare our outputs to something else. Nevertheless, we provided the SARI scores and examples of simplification to demonstrate our models' performance. The metrics and the examples indicated that optimal simplification performance can, in fact, be achieved under the conditions of our experiments. Not only training neural networks from scratch, but also utilizing pre-trained language models such as mBART and fine-tuning them on the created datasets, proved effective for simplifying both Russian and Finnish. The morphological richness of these languages, however, might have been the reason for some of the models' hallucinations and/or ungrammatical outputs. It should be noted, again, that we never aimed to select the optimal model parameters for the tasks at hand, but rather to prove that a certain setup worked (for example, that a new dataset could be successfully used for training simplification models). Moreover, with the introduction of efficient fine-tuning techniques for larger models, the problem of suboptimal outputs for sequence-to-sequence tasks even with little available data seems to be more mitigable.

We have also proven that transfer learning strategies can work under low-resource conditions by experimenting with English data in Paper I. In that paper, we achieved the best results when jointly training on both the data-rich task (summarization) and the data-poor task (simplification). The experiments in Paper V demonstrate that data augmentation with data from similar domains (such as selecting paraphrases where one of the texts is simpler than the other to augment simplification data) can aid in training for more complex tasks, such as controllable simplification. This paper also supports the claim that joint learning can be beneficial, but only for some sets of tasks: training simultaneously for controlling both the amount of compression and the amount of paraphrasing was successful, while adding the task of controlling the CEFR level to this set resulted in decreasing performance.

At the moment, it is hard to outline concrete directions for **future work** in the field of automatic text simplification. Natural language processing is a rapidly evolving field, which means that new data sources, methods, and models are constantly being developed. While working on this thesis project, we always tried to compare multiple approaches to find the best solution at the moment.

On the one hand, it feels exciting to imagine that simplification tools will soon be as accessible and as widely used as machine translation tools. It can already be seen that with increased model size comes increased simplification quality (see Vadlamannati and Şahin, 2023) for a comparison of LLMs of different sizes' simplification performance); therefore, the users of large language model applications have the opportunity to obtain good quality simplifications of their desired texts, provided that the text fits the model's context window. Therefore, it

is reasonable to assume that in the future researchers will train larger models for simplification in order to overcome shortcomings such as incoherent and ungrammatical outputs. With the increasing availability of larger models, the simplification algorithms evolve, too: for example, today one can use prompt engineering and in-context learning instead of fine-tuning to perform controlled simplification.

On the other hand, certain problems come with this rapid development, namely the environmental impact of running large language models and the cost of maintaining such services, which makes providing completely free and unlimited NLP applications almost impossible. Therefore, we can suppose that researchers will continue to experiment with using smaller models for simplification. For example, it has been shown that a controllable lexical simplification system based on the T5 architecture can outperform a system based on text-davinci-002, a model with 176B parameters, on several evaluation metrics (Kim and Saggion, 2023; metrics described in Saggion et al. (2022)).

The growing body of research on simplification models will facilitate the need for more sophisticated automatic evaluation. For example, this thesis project's limitation of only using automated evaluation metrics could have been partially mitigated by utilizing a learnable metric such as LENS (Maddela et al., 2023) or REFeREE (Huang and Kochmar, 2024) – measures that employ models trained on annotated data to mimic human judgments of simplicity. Because of their data-driven nature, these metrics correlate better with human judgments than non-learnable metrics such as SARI. These approaches, however, are not language agnostic and are not available for languages such as Russian or Finnish. Hopefully, the number of learnable simplification evaluation metrics for different languages will increase in the near future.

Despite the rapid growth, it seems like there is still not enough data for even the biggest models to cover all the different domains and aspects of accessible communication. While working on this project, I attempted to bridge this gap by making simplification datasets for languages that, at that moment, did not have such resources. It has been demonstrated that these data can be used for successfully training simplification models of various kinds, and these datasets are now openly available for non-commercial use. The future of automatic simplification hopefully lies in making everything that we as researchers invent and produce as accessible as possible in every sense: more open-access materials, more languages and audiences covered, more domains explored beyond text. My personal future work plans include maintenance of the concluded projects (such as the parallel datasets that were created over the course of working on this thesis) and continuing the work on simplification models, increasing their quality while keeping the use of resources moderate.

# References


Alonzo, O., Seita, M., Glasser, A., & Huenerfauth, M. (2020). Automatic Text Simplification Tools for Deaf and Hard of Hearing Adults: Benefits of Lexical Simplification and Providing Users with Autonomy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3313831.3376563

Aluísio, S., & Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In T. Solorio & T. Pedersen (Eds.), Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (pp. 46–53). Association for Computational Linguistics. https://aclanthology.org/W10-1607

Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C., & Specia, L. (2017). Learning How to Simplify From Explicit Labeling of Complex-Simplified Text Pairs. In G. Kondrak & T. Watanabe (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 295–305). Asian Federation of Natural Language Processing. https://aclanthology.org/I17-1030

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., & Specia, L. (2020). ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4668–4679). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.424

Alva-Manchego, F., Martin, L., Scarton, C., & Specia, L. (2019). EASSE: Easier Automatic Sentence Simplification Evaluation. In S. Padó & R. Huang (Eds.), 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 (pp. 49–54). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3009

Anastasyev, D. (2021). RuSimpleSentEval. https://github.com/DanAnastasyev/RuSimpleSentEval

Bouillon, P., Gerlach, J., Mutal, J., Tsourakis, N., & Spechbach, H. (2021). A Speech-enabled Fixed-phrase Translator for Healthcare Accessibility. In A. Field, S. Prabhumoye, M. Sap, Z. Jin, J. Zhao, & C. Brockett (Eds.), Proceedings of the 1st Workshop on NLP for Positive Impact (pp. 135–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.15

Bredel, U., & Maaß, C. (2017). Wortverstehen durch Wortgliederung–Bindestrich und Mediopunkt in Leichter Sprache. In B. M. Bock, U. Fix, & D. Lange

(Eds.), “Leichte Sprache” im Spiegel theoretischer und angewandter Forschung (pp. 211–228). Berlin: Frank & Timme.

Brouwers, L., Bernhard, D., Ligozat, A.-L., & François, T. (2014). Syntactic sentence simplification for French. Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL 2014, 47–56.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners.

Chandrasekar, R., & Bangalore, S. (1997). Automatic induction of rules for text simplification. Knowledge-Based Computer Systems: Research and Applications.

Chandrasekar, R., Doran, C., & Bangalore, S. (1996). Motivations and methods for text simplification. COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics.

Cooper, M., & Shardlow, M. (2020). CombiNMT: An Exploration into Neural Text Simplification Models. Proceedings of the 12th Language Resources and Evaluation Conference, 5588–5594. https://www.aclweb.org/anthology/2020.lrec-1.686

Coster, W., & Kauchak, D. (2011). Simple English Wikipedia: A New Text Simplification Task. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 665–669). Association for Computational Linguistics. https://aclanthology.org/P11-2117

Council of Europe, Council for Cultural Co-operation, Education Committee, & Modern Languages Division. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Creutz, M. (2018). Open Subtitles Paraphrase Corpus for Six Languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA). https://aclanthology.org/L18-1218

Cripwell, L., Legrand, J., & Gardent, C. (2023). Context-Aware Document Simplification. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13190–13206). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.834

De Belder, J., & Moens, M.-F. (2010). Text simplification for children. Prroceedings of the SIGIR Workshop on Accessible Search Systems, 19–26.

Degraeuwe, J., & Saggion, H. (2022). Lexical Simplification in Foreign Language Learning: Creating Pedagogically Suitable Simplified Example Sentences. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, & W. Xu (Eds.), Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022) (pp. 98–110). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.tsar-1.9

Deilen, S. (2022). Visual segmentation of compounds in Easy Language: does the marking of morpheme boundaries reduce cognitive processing costs? In M. P. Castillo Bernal & M. Estévez Grossi (Eds.), Translation, Mediation and Accessibility for Linguistic Minorities (TransÜD: Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens) (pp. 161–174). Berlin: Frank & Timme.

Devaraj, A., Marshall, I., Wallace, B., & Li, J. J. (2021). Paragraph-level Simplification of Medical Texts. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4972–4984). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.395

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

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

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

Dong, Y., Li, Z., Rezagholizadeh, M., & Cheung, J. C. K. (2019). EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3393–3402). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1331

Dou, Z.-Y., & Neubig, G. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2112–2128.

Dras, M. (1999). Tree adjoining grammar and the reluctant paraphrasing of text [Phdthesis]. Macquarie University, Australia.

Elior, S., Omri, A., & Ari, R. (2018). BLEU is Not Suitable for the Evaluation of Text Simplification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 738–744. https://doi.org/10.18653/v1/d18-1081

Espinosa-Zaragoza, I., Abreu-Salas, J., Moreda, P., & Palomar, M. (2023). Automatic Text Simplification for People with Cognitive Disabilities: Resource Creation within the ClearText Project. In S. Štajner, H. Saggio, M. Shardlow, & F. Alva-Manchego (Eds.), Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability (pp. 68–77). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.tsar-1.7

European Commission. (2015). How to write clearly. Directorate-General for Translation. https://doi.org/10.2782/022405

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J. C. (2020). Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 1353–1361). European Language Resources Association. https://aclanthology.org/2020.lrec-1.169

Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., & Kambhatla, N. (2022). Text Simplification for Legal Domain: Insights and Challenges. In N. Aletras, I. Chalkidis, L. Barrett, C. Goanullță, & D. Preonulltiuc-Pietro (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2022 (pp. 296–304). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nllp-1.28

Gehrmann, S., Deng, Y., & Rush, A. (2018). Bottom-Up Abstractive Summarization. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4098–4109). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1443

González-Sordé, M., & Matamala, A. (2023). Empirical evaluation of Easy Language recommendations: a systematic literature review from journal research in Catalan, English, and Spanish. Universal Access in the Information Society. https://doi.org/10.1007/s10209-023-00975-2

Gudkov, V., Mitrofanova, O., & Filippskikh, E. (2020). Automatically Ranked Russian Paraphrase Corpus for Text Generation. In A. Birch, A. Finch, H. Hayashi, K. Heafield, M. Junczys-Dowmunt, I. Konstas, X. Li, G. Neubig, & Y. Oda (Eds.), Proceedings of the Fourth Workshop on Neural Generation and Translation (pp. 54–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.ngt-1.6

Herget, K., & Alegre, T. (2023). Simplification Strategies in the Field of Environment and Climate Change: Exploring ChatGPT for Scientific Popularisation in LSP Classes. INTERNATIONAL CONFERENCE PROCEEDINGS - INNOVATION IN LANGUAGE LEARNING 16th Edition (9-10 November 2023 | in Florence and Online). https://doi.org/https://doi.org/10.26352/HY09_2384-9509

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, Y., & Kochmar, E. (2024). REFeREE: A REference-FREE Model-Based Metric for Text Simplification. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 13740–13753). ELRA. https://aclanthology.org/2024.lrec-main.1200/

Inclusion Europe. (2010). European standards for making information easy to read and understand. Pathways to adult education for people with intellectual disabilities. https://www.inclusion-europe.eu/wp-content/uploads/2017/06/EN_Information_for_all.pdf

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J., & others. (2023). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. European Radiology. https://doi.org/10.1007/s00330-023-10213-1

Katsuta, A., & Yamamoto, K. (2018). Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation(LREC 2018). European Language Resources Association (ELRA). https://aclanthology.org/L18-1072

Kauchak, D. (2013). Improving Text Simplification Language Modeling Using Unsimplified Text Data. In H. Schuetze, P. Fung, & M. Poesio (Eds.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1537–1546). Association for Computational Linguistics. https://aclanthology.org/P13-1151

Khallaf, N., Sharoff, S., & Soliman, R. (2022). Towards Arabic Sentence Simplification via Classification and Generative Approaches. In H. Bouamor, H. Al-Khalifa, K. Darwish, O. Rambow, F. Bougares, A. Abdelali, N. Tomeh, S. Khalifa, & W. Zaghouani (Eds.), Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP) (pp. 43–52). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wanlp-1.5

Khitriuk, V. V., Soroko, J. N., Grishan, T. V., Kovaleva, V. I., & Braun, B. (2018). «Ясный язык»: как сделать информацию доступной для чтения и понимания. Методические рекомендации (J. G. Titova, Ed.). https://lifeguide.by/wp-content/uploads/2021/10/metodicheskie-rekomendaczii-yasnyj-yazyk.pdf

Kim, S. C., & Saggion, H. (2023). Multilingual Controllable Transformer-Based Lexical Simplification. Proces. Del Leng. Natural, 71, 109–123. https://api.semanticscholar.org/CorpusID:259344369

Kincaid, J. P., Fishburne, R., Rogers, R., & Chissom, B. (1975). Derivation of new readability formulas (Automated Reliability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Naval Technical Training, US Naval Air Station: Millington, TN.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of ACL 2017, System Demonstrations, 67–72. https://www.aclweb.org/anthology/P17-4012

Kulkki-Nieminen, A. (2010). Selkoistettu uutinen. Lingvistinen analyysi selkotekstin erityispiirteistä [Phdthesis, Tampere University Press]. https://urn.fi/urn:isbn:978-951-44-8093-5

Lal, P., & Ruger, S. (2002). Extract-based summarization with simplification. Proceedings of the ACL.

Laposhina, A., Veselovskaya, T., & Krivenko, O. (2019, June). Иллюстративно-текстовый корпус учебников русского языка для детей младшего школьного возраста: концепция и методика создания. Труды Международной Конференции “Корпусная Лингвистика - 2019.”

Laposhina, A., Veselovskaya, T., Lebedeva, M., & Krivenko, O. (2018). Automated Text Readability Assessment For Russian Second Language Learners. Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies “DIALOGUE,” 396–406. https://www.dialog-21.ru/media/5216/laposhina%D0%B0nplusetal.pdf

Leskelä, L. (2021). Easy Language in Finland. In C. Lindholm & U. Vanhatalo (Eds.), Handbook of Easy Languages in Europe (1st ed., Vol. 8, pp. 149–190). Berlin: Frank & Timme. https://doi.org/10.26530/20.500.12657/52628

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Li, Z., Shardlow, M., & Alva-Manchego, F. (2023). Comparing Generic and Expert Models for Genre-Specific Text Simplification. Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, 51–67.

Lindholm, C., & Vanhatalo, U. (2021). Introduction. In C. Lindholm & U. Vanhatalo (Eds.), Handbook of Easy Languages in Europe (1st ed., Vol. 8, pp. 11–26). Frank & Timme. https://doi.org/10.26530/20.500.12657/52628

Liu, L., & Zhu, M. (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. https://doi.org/10.1093/llc/fqac089

Liu, X. (2024). A Survey of Hallucination Problems Based on Large Language Models. Proceedings of the 2nd International Conference on Machine Learning and Automation, 97, 24–30. https://doi.org/10.54254/2755-2721/97/2024.17851

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics, 8, 726–742. https://doi.org/10.1162/tacl_a_00343

Luukkonen, R., Komulainen, V., Luoma, J., Eskelinen, A., Kanerva, J., Kupari, H.-M., Ginter, F., Laippala, V., Muennighoff, N., Piktus, A., Wang, T., Tazi, N., Scao, T., Wolf, T., Suominen, O., Sairanen, S., Merioksa, M., Heinonen, J., Vahtola, A., … Pyysalo, S. (2023). FinGPT: Large Generative Models for a Small Language. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 2710–2726). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.164

Lyu, Q., Tan, J., Zapadka, M. E., Ponnatapura, J., Niu, C., Myers, K. J., Wang, G., & Whitlow, C. T. (2023). Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Visual Computing for Industry, Biomedicine, and Art, 6(1), 9.

Maaß, C. (2020). Easy Language – Plain Language – Easy Language Plus (1st ed., Vol. 3). Frank & Timme. https://www.frank-timme.de/en/programme/product/easy_language-plain_language-easy_language_plus

Maddela, M., Dou, Y., Heineman, D., & Xu, W. (2023). LENS: A Learnable Evaluation Metric for Text Simplification. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16383–16408). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.905

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., & Sagot, B. (2022). MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 1651–1664). European Language Resources Association. https://aclanthology.org/2022.lrec-1.176

Martin, L., Villemonte de La Clergerie, É., Sagot, B., & Bordes, A. (2020). Controllable Sentence Simplification. LREC 2020 - 12th Language Resources and Evaluation Conference. https://hal.inria.fr/hal-02678214

Maruyama, T., & Yamamoto, K. (2018). Simplified Corpus with Core Vocabulary. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA). https://aclanthology.org/L18-1185

Matamala, A. (2016). The ALST Project: Technologies for Audio Description. In A. Matamala & P. Orero (Eds.), Researching Audio Description: New Approaches (pp. 269–284). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-56917-2_14

Moreno-Sandoval, A., Campillos-Llanos, L., & García-Serrano, A. (2024). Language Resources in Spanish for Automatic Text Simplification across Domains. https://arxiv.org/abs/2409.20466

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., & Xiang, B. (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In S. Riezler & Y. Goldberg (Eds.), Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (pp. 280–290). Association for Computational Linguistics. https://doi.org/10.18653/v1/K16-1028

Nechaeva, N., Helmle, K.-S., & Kairova, E. (2020). Перевод на ясный и простой языки: зарубежный опыт и перспективы в России. In Вестник ПНИПУ. Проблемы языкознания и педагогики (Vol. 3).

Netzwerk Leichte Sprache. (2022). Die Regeln für Leichte Sprache. https://www.leichte-sprache.org/wp-content/uploads/2023/03/Regelwerk_NLS_Neuaufl2022_web.pdf

Nishihara, D., Kajiwara, T., & Arase, Y. (2019). Controllable Text Simplification with Lexical Constraint Loss. In F. Alva-Manchego, E. Choi, & D. Khashabi (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 260–266). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2036

Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017). Exploring Neural Text Simplification Models. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 85–91. https://doi.org/10.18653/v1/P17-2014

Ogden, C. K. (1940). General Basic English Dictionary. Evans Brothers Limited. https://books.google.lt/books?id=8GyE0AEACAAJ

Östling, R., & Tiedemann, J. (2016). Efficient word alignment with Markov Chain Monte Carlo. Prague Bulletin of Mathematical Linguistics, 106, 125–146. http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Proceedings of NAACL-HLT 2019: Demonstrations.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback.

Paetzold, G., Alva-Manchego, F., & Specia, L. (2017). MASSAlign: Alignment and Annotation of Comparable Documents. Proceedings of the IJCNLP 2017, System Demonstrations, 1–4. https://aclanthology.org/I17-3001

Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. Journal of Artificial Intelligence Research, 60, 549–593.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

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

Perego, E. (2020). Accessible Communication: A Cross-country Journey (1st ed., Vol. 4). Frank & Timme. https://www.frank-timme.de/de/programm/produkt/accessible%5C_communication-a%5C_cross-country%5C_journey

Pereira, F. V., Frazão, A., & Moreira, V. P. (2024). Automatic Text Simplification for the Legal Domain in Brazilian Portuguese. Brazilian Conference on Intelligent Systems. https://api.semanticscholar.org/CorpusID:276618652

Qiao, Y., Li, X., Wiechmann, D., & Kerz, E. (2022). (Psycho-)Linguistic Features Meet Transformer Models for Improved Explainable and Controllable Text Simplification. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, & W. Xu (Eds.), Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022) (pp. 125–146). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.tsar-1.12

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Ruel, J., Allaire, C., Moreau, A. C., Kassi, B., Brumagne, A., Delample, A., Grisard, C., & Pinto da Silva, F. (2018). Communiquer pour tous. Guide pour une information accessible (C. Allaire, Ed.). Saint-Maurice : Santé publique France. https://www.cnsa.fr/documentation/ns04-112-18l_spf_communiquer_pour_tous_bd_total_web.pdf

Ryan, M., Naous, T., & Xu, W. (2023). Revisiting non-English Text Simplification: A Unified Multilingual Benchmark. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 4898–4927).

Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.269

Saggion, H. (2017). Automatic Text Simplification. In Synthesis Lectures on Human Language Technologies. Springer International Publishing. https://doi.org/10.1007/978-3-031-02166-4

Saggion, H., Štajner, S., Ferrés, D., Sheang, K. C., Shardlow, M., North, K., & Zampieri, M. (2022). Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, & W. Xu (Eds.), Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022) (pp. 271–283). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.tsar-1.31

Sakhovskiy, A., Izhevskaya, A., Pestova, A., Tutubalina, E., Malykh, V., Smurov, I., & Artemova, E. (2021). RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian. 607–617. https://doi.org/10.28995/2075-7182-2021-20-607-617

Scarton, C., Aprosio, A. P., Tonelli, S., Wanton, T. M., & Specia, L. (2017). MUSST: A multilingual syntactic simplification tool. Proceedings of the IJCNLP 2017, System Demonstrations, 25–28.

Scarton, C., & Specia, L. (2018). Learning Simplifications for Specific Target Audiences. In I. Gurevych & Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 712–718). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2113

Schioppa, A., Vilar, D., Sokolov, A., & Filippova, K. (2021). Controlling Machine Translation for Multiple Attributes with Additive Interventions. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6676–6696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.535

See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1073–1083). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1099

Sennrich, R., & Volk, M. (2010). MT-based Sentence Alignment for OCR-generated Parallel Texts. Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers. https://aclanthology.org/2010.amta-papers.14

Shterionov, D., Sisto, M. D., Muller, M., Landuyt, D. V., Omardeen, R., Oboyle, S., Braffort, A., Roelofsen, F., Blain, F., Vanroy, B., & Avramidis, E. (Eds.). (2023). Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages. European Association for Machine Translation. https://aclanthology.org/2023.at4ssl-1.0

Siddharthan, A. (2014). A survey of research on text simplification. ITL-International Journal of Applied Linguistics, 165(2), 259–298.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 16857–16867). Curran Associates, Inc.

https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

Spring, N., Kostrzewa, M., Fröhlich, D., Rios, A., Pfütze, D., Battisti, A., & Ebling, S. (2023). Analyzing sentence alignment for automatic simplification of German texts. In S. Deilen, S. Hansen-Schirra, S. H. Garrido, C. Maaß, & A. Tardel (Eds.), Emerging Fields in Easy Language and Accessible Communication Research (pp. 339–369). Frank & Timme GmbH. https://doi.org/10.57088/978-3-7329-9026-9_13

Štajner, S., Franco-Salvador, M., Ponzetto, S. P., Rosso, P., & Stuckenschmidt, H. (2017). Sentence Alignment Methods for Improving Text Simplification Systems. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 97–102). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2016

Stodden, R., Momen, O., & Kallmeyer, L. (2023). DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16441–16463). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.908

Taaltelefoon. (2012). In duidelijk Nederlands. Spreken en schrijven voor iedereen (D. Caluwé, A. Bosmans, S. Croon, S. De Schepper, K. Maesen, K. Spillebeen, S. Van Calster, & V. Verreycken, Eds.). Vlaamse overheid. https://assets.vlaanderen.be/image/upload/v1645443512/In_duidelijk_Nederlands_139_xepeac.pdf

Tanprasert, T., & Kauchak, D. (2021). Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, & W. Xu (Eds.), Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021) (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.1

Thompson, B., & Koehn, P. (2020). Exploiting Sentence Order in Document Alignment. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5997–6007. https://doi.org/10.18653/v1/2020.emnlp-main.483

Thompson, B., & Koehn, P. (2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 1342–1348. https://doi.org/10.18653/v1/D19-1136

Tonelli, S., Aprosio, A. P., & Saltori, F. (2016). SIMPITIKI: a Simplification corpus for Italian. CLiC-It/EVALITA. https://api.semanticscholar.org/CorpusID:11914765

Vadlamannati, S., & Şahin, G. (2023). Metric-Based In-context Learning: A Case Study in Text Simplification. In C. M. Keet, H.-Y. Lee, & S. Zarrieß (Eds.), Proceedings of the 16th International Natural Language Generation Conference (pp. 253–268). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.inlg-main.18

Vandeghinste, V., Schuurman, I., Sevens, L., & Van Eynde, F. (2017). Translating text into pictographs. Natural Language Engineering, 23(2), 217–244. https://doi.org/10.1017/S135132491500039X

Vaschalde, C., Trial, P., Esperança-Rodier, E., Schwab, D., & Lecouteux, B. (2018, November). Automatic pictogram generation from speech to help the implementation of a mediated communication. Conference on Barrier-Free Communication. https://hal.science/hal-01880744

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. Advances in Neural Information Processing Systems, 28.

Voita, E. (2020). NLP Course For You. https://lena-voita.github.io/nlp_course.html

Wang, T., Chen, P., Rochford, J., & Qiang, J. (2016). Text Simplification Using Neural Machine Translation. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). https://doi.org/10.1609/aaai.v30i1.9933

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. Transactions of the Association for Computational Linguistics, 3, 283–297. https://doi.org/10.1162/tacl_a_00139

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016a). Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics, 4, 401–415. https://doi.org/10.1162/tacl_a_00107

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 483–498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2020). Multilingual Universal Sentence Encoder for Semantic Retrieval. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12

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

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

Yleisradio. (2021a). Yle Finnish News Archive 2019-2020, source. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2021050402

Yleisradio. (2021b). Yle News Archive Easy-to-read Finnish 2019-2020, source. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2021050702

Zhang, X., & Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. In M. Palmer, R. Hwa, & S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 584–594). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1062

Абрамов, А. В. and Иванов, В. В. (2022). Сбор и оценка лексической сложности данных для русского языка с помощью краудсорсинга. Russian Journal of Linguistics, 26(2), 409–425. https://journals.rudn.ru/linguistics/article/view/31331

Андрюшина, Н. П., Битехтина, Г. А., Клобукова, Л. П., Норейко, Л. Н., & Одинцова, И. В. (2019). Лексический минимум по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. ООО Центр “Златоуст.”

Дале, Д. (2021). Перефразирование русских текстов: корпуса, модели, метрики. https://habr.com/ru/post/564916/

Ляшевская, Ольга and Шаров, Сергей. (2009). Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). Москва: Азбуковник.

Ростова, Е. Г. (2018). Мультимедийный лингвострановедческий словарь “Россия” — в помощь изучающим русский язык и культуру в СНГ. Вестник Библиотечной Ассамблеи Евразии, 1, 46–48.