Malmon: A Crowd-Sourcing Platform for Simple Language

Helgi Björn Hjartarson, Steinunn Rut Friðriksdóttir
University of Iceland
Sæmundargata 2, Reykjavík
hbh42@hi.is, srf2@hi.is

Abstract

This paper presents a crowd-sourcing platform designed to address the need for parallel corpora in the field of Automatic Text Simplification (ATS). ATS aims to automatically reduce the linguistic complexity of text to aid individuals with reading difficulties, such as those with cognitive disorders, dyslexia, children, and non-native speakers. ATS does not only facilitate improved reading comprehension among these groups but can also enhance the preprocessing stage for various NLP tasks through summarization, contextual simplification, and paraphrasing. Our work introduces a language independent, openly accessible platform that crowdsources training data for ATS models, potentially benefiting low-resource languages where parallel data is scarce. The platform can efficiently aid in the collection of parallel corpora by providing a user-friendly data-collection environment. Furthermore, using human crowd-workers for the data collection process offers a potential resource for linguistic research on text simplification practices. The paper discusses the platform's architecture, built with modern web technologies, and its user-friendly interface designed to encourage widespread participation. Through gamification and a robust admin panel, the platform incentivizes high-quality data collection and engagement from crowdworkers.

Keywords: automatic text simplification, crowd-sourcing platform, gamification

1. Introduction

Automatic text simplification (ATS) is a Natural Language Processing (NLP) task where the linguistic complexity of text is reduced in order to facilitate reading comprehension without losing its original information. This is particularly helpful for readers with low literacy, for instance due to cognitive disorders or dyslexia, children and non-native speakers learning a new language. Text simplification has also been shown to improve results when used at the preprocessing stage for other NLP tasks. The ATS process varies in nature; for instance, it can involve summarizing the text to remove any redundant information, simplifying the context of the text, or paraphrasing it so that key points are emphasized. Usually, ATS involves two steps, lexical simplification and syntactic simplification, where the former focuses on reducing complexity by replacing complex words with simpler synonyms and the latter reduces grammatical complexity, such as by removing or simplifying subordinate clauses that may be difficult for readers to comprehend.

One of the challenges of ATS is identifying the complexity of a given text and deciding the best way to reduce it. In order to train models that can perform this task automatically, it is fundamental to have access to extensive parallel corpora in which complex sentences are paired with their simplified versions. Various ATS corpora exist for high-resource languages like English (see Al-Thanyyan and Azmi (2021) for an overview) but far fewer for lower-resource languages. Recent methodologies in ATS are largely data-driven where simplification rules are inducted from the data. It is therefore crucial to create simple, ready-to-use tools that researchers can use for their ATS data collection. In our work, we introduce Malmon, a crowdsourcing platform that can be used to collect training data for text simplification models. The platform is language independent, openly accessible[1] and easily adaptable for researchers wanting to collect their own data.

2. Collecting ATS Data

As creating an ATS corpus from scratch can be prohibitively expensive, many attempts at automating the data collection have been made. In their paper, Holmer and Rennes (2023) describe the creation of a pseudo-parallel ATS corpus for Swedish. They then fine-tune a BART model for sentence simplification on their data with promising results. Ormaechea and Tsourakis (2023) use a combination of automatic methods and manual annotation to align Wikipedia articles in French to their counterparts in the simplified version Vikidia. Similarly, Dmitrieva and Konovalova (2023) use Sentence Transformers to measure the similarity between Finnish news articles and their simplified counterparts to create sentence pairs which are then manually reviewed.

These semi-automatic methods can potentially speed up the data collection process and consequently reduce the cost required. However, they

[1]The source code for the platform is available on https://github.com/polarparsnip/malmon. We encourage other researchers to use and adapt this code to their needs.

3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI), pages 15–21
20-25 May, 2024May 20th, 2024.

are not without their setbacks. Holmer and Rennes (2023) mention some problems relating to the automating process, for instance that sentences with named entities often get aligned with sentences containing completely different entities. Another approach is to have expert annotators manually simplify sentences or excerpts of text. This approach was used for the Newsela corpus, whose 2016-01-29.1 version consists of 1,911 news articles and up to 5 simplified versions written by trained professionals (Paetzold and Specia, 2017). Similarly, the Alector corpus, intended to research the effectiveness of simplifying text for dyslexic children, was constructed by a group of experts who manually simplified 79 literary and scientific texts commonly used in French schools (Gala et al., 2020).

A similar approach, and the one we advocate here, is to collect ATS data using crowdsourcing. Katsuta and Yamamoto (2018) crowdsourced a parallel corpus for Japanese. Crowdworkers were asked to limit themselves to a core vocabulary of 2000 words so that the resulting simplifications was at an everyday conversational level. Appendix A shows the proposed guidelines for crowdworkers using our platform to simplify Icelandic text. While our guidelines do not include a core vocabulary similar to that of Katsuta and Yamamoto (2018), we encourage our users to avoid rare or complex words. We note that this frame of reference can be changed at will so that it better suits the needs of other researchers using the platform. We also note that our platform can be used both for crowdsourcing simplified versions of text examples directly (whether the data collection process is to be open to the public or conducted by expert crowdworkers) and for manually reviewing sentence alignments created with automatic methods. If the former is chosen, the resulting data could also be used in linguistic research on the way people simplify or paraphrase complex text, similarly to what was presented in Amancio and Specia (2014).

3. The Platform

3.1. Motivation

As previously mentioned, ATS can greatly benefit individuals with low literacy levels. Azab et al. (2015) used ATS methods to design a browser extension to help students learning English as a second language by annotating and substituting difficult words with simpler synonyms. A similar platform for people with aphasia was designed by Devlin and Unthank (2006), which presents users who have difficulty understanding or remembering a particular word with another word that has the same meaning but is more common or easier to understand. Javourey-Drevet et al. (2022) designed an iPad application that presented French children with original and simplified versions of informative and narrative texts. Their results indicate that the simplified texts benefited poor readers and children with weaker cognitive skills, increasing their reading fluency and text comprehension.

ATS methods can also facilitate comprehension of particularly difficult or domain-specific text. This has been prominent within the medical field. Kushalnagar et al. (2018) used ATS methods to simplify information about breast cancer in order to improve comprehension for Deaf people who use American Sign Language. Phatak (2023) proposed several ATS methods to simplify complex biomedical literature in English for the general public and Cardon and Grabar (2020) did the same for French. Similarly, Truică et al. (2023) presented SimpLex, a software that uses ATS methods to simplify medical text in English for the general public. ATS systems have also been used as a preprocessing task for other NLP systems to improve their results. In their paper, Van et al. (2021) show that augmenting data with ATS to provide additional information during training significantly improves performances of various text classification and relation extraction models.

However, to be able to create such systems, it is fundamental that sufficient parallel data exist. TS-ANNO, introduced by Stodden and Kallmeyer (2022), is a crowd-sourcing platform that can be used for a variety of tasks related to ATS. Our platform, however, focuses solely on generating parallel complex-simple sentence pairs. It offers a simple, easy to use way of collecting the data via crowdsourcing. As discussed in Section 3.3, the users of our platform are presented with three options only, to simplify, verify or download the resulting data. This straight-forward navigation leaves little room for confusion as to what is expected of the users, particularly crowdworkers that might not have previous experience with work in NLP. Our platform may prove especially useful for lower-resource languages where the number of expert annotators might be scarce and priority must be placed on straight-forward solutions aimed at the general public. Researchers interested in using the platform can access the source code and modify it freely.

3.2. Technical Information

The platform, which we call Malmon, is built as a full-stack website utilizing a SQL database set up with PostgreSQL to store all sentences and user data. The back-end web server is built in Javascript utilizing Express.js to handle http connections and the front-end of the website is made using Next.js, which is a React-based Javascript framework. When a user is logged in, the server

checks if they are an admin or a general user and redirects them to the appropriate section of the site. The server makes sure that general users can't access any of the admin areas and that a logged in admin has access to all necessary admin functionalities.

User registration and login on the site are straightforward. Users are required to enter a username, e-mail, and password when they register an account. Since user accounts are tied to each individual user's progress, it is important to be able to recover an account in the event that a user loses their password. Tying e-mail addresses to user accounts could also possibly help to distinguish between different users if the need arises, for instance to detect outliers that may be the result of system spamming.

The language of the platform can be chosen with one environment variable when setting it up for hosting, with current supported languages being: Icelandic, English, Norwegian, Danish, Swedish, Faroese, and Italian[2]. Additionally, adjusting existing language settings or adding more supported languages is a straightforward process that only involves modifying one file.

3.3. Functionality

When not logged in, site visitors are presented with a simple website with a front page detailing how the platform works. In the footer, they have the option to log in or to register a new account. In the navigation menu, they again have two options: to log in and to get data. The latter option is the only functionality available to users when they are not logged in, apart from actions such as creating a new account or signing in. This option allows visitors to the platform to fetch the current state of the resulting dataset as either a JSON or CSV file, with the files containing complex-simplified sentence pairs. This means that the dataset being collected at each given time is open to everyone who wishes to use it for model training or other similar purposes.

Once logged in, users still see the option to download the dataset but are also presented with options in the navigation menu that are only visible to logged in users. They now see options for navigating to their *account* section and a *score-table* section, both of which will be covered in more detail in Section 3.4. They are also presented with the option to go to an FAQ page detailing the guidelines for submitting sentences and the options to go to the *simplify* (see Figure 1) and the *verify* sections (see Figure 2). These two last sections are the main areas in which users contribute to the dataset being collected on the platform.

Once users navigate to the *simplify* section, they are given a random sentence from the database and a fast and simple CAPTCHA-like task to verify that they are a human and not a bot. After completing the task, they are presented with an input text field in which they can enter a simplified version of the sentence they were given. This CAPTCHA-like task makes sure sentences are being submitted by humans and prevents bots and/or other spam methods from being able to submit sentences. Once the user feels like they have entered a good enough simplified version of the sentence they received, they can click submit and the sentence is then saved in the database.

When a simplified sentence submitted by a user is saved in the database, it is marked as unverified and is therefore not yet part of the dataset which can be downloaded. To be included in the dataset, a submitted sentence must first be verified by a separate user on the platform which is the purpose of the *verify* section. Once users navigate to that section, they are again given a sentence along with a simplified version of that sentence submitted by another user. After completing the same CAPTCHA-like task, they are presented with two buttons, a "confirm" button and a "reject" button. If a user feels like the simplification submitted by another user is a good representation of the original sentence, they can approve it by pressing the "confirm" button. If they feel like the simplified version is not a good representation of the original sentence, they can reject it by pressing the "reject" button. If a simplified sentence is confirmed by the user, it is marked as *verified* and is now part of the collected dataset which people can download. If it is rejected, it is taken out of circulation and will no longer appear to users.

These two sections form the data collection portion of the platform and are the main ways users interact with the website.

3.4. Gamification

Crowdsourcing is a data collection process whereby content is obtained by having a group of people use their leisure time to make their contributions at a minimal cost. As crowdsourcing is generally performed by non-expert volunteers, there needs to be some incentive for participation, as what is considered interesting from a scientific perspective may not be enjoyable for the general public. One way to achieve this is through gamification, which incorporates video game elements to improve user experience in a non-game service which in turn can enhance user engagement (see for instance Deterding et al., 2011; Quecke and Mariani, 2021). Competitive game elements, such

[2]Note that the proposed guidelines are only available in Icelandic and English as of this publication. The other languages have been translated using ChatGPT and thus require further review.

Figure 1: The simplification page of the platform. Users are presented with a complex sentence and are asked to write a simplified version of the sentence.

as points and immediate performance feedback, have been found to positively affect crowdworker motivation and, consequently, participation (Yang et al., 2021).

In our site's navigation menu, all users can access a *score-table* that details which users have contributed the most in terms of submitted simplified sentences and the amount of submitted sentence verifications. Users are ranked based on the lower of the two aforementioned attributes, so if a user has, for example, submitted 33 simplified sentences and verified 22 sentences, they will be ranked based on the number 22. This guarantees that users can't focus exclusively on one method in order to receive a good score, instead providing incentive to contribute to both areas in order to boost their ranking on the scoreboard.

In the *account* section, users can view their information which includes their username, as well as how many sentences they have submitted and how many sentence verifications they have completed. Also contained in the *account* section is a digital pet tied to their account that grows according to their sentence submissions and sentence verifications, based on the same system as the scoreboard. When a user has only just created an account and not yet taken any action, the digital pet appears as an egg. As they contribute to the platform, their pet evolves into higher stages similar to creatures in franchises like Pokémon or Digimon.

These features encourage and reward users for their contributions to the platform and can act as a basis for other reward systems which could then be integrated with them. For instance, the fully evolved pet could be accompanied by a lottery ticket in the form of a QR code where a diligent user gains the chance to win a real-world price. Adding an image to the last stage of the pet is straightforward and only involves adding two environment variables when the platform is set up in hosting. Since users have to confirm they are human before submitting sentences, it will be difficult to try to cheat the system to gain whatever rewards are in place.

3.5. Admin Functionality

One of the key focus points for the platform was to have extensive and user friendly admin functionality. When an admin is logged in, they have instant access to the admin dashboard. This dashboard allows an admin to access the editor areas for sentences, simplified sentences, and users.

In the *sentences* area, an admin can view all the saved sentences from the database. The sentences are displayed 10 at a time with the option to move forward to the next 10 sentences. Each sentence is displayed individually with an option to update that specific sentence or delete it, so if an admin notices a sentence containing errors or one that should not be there, they have the option to react accordingly. There is also a form on the page where an admin can register a new sentence and add it to the list of complex sentences.

In the *simplified sentences* area, an admin can view all the simplified sentences that have been submitted, 10 at a time. The simplified sentences are displayed individually with information on whether the sentence has been confirmed or rejected by another user. They are also accompanied by options for an admin to either delete the sentence or delete a user rejection. An admin may delete a user rejection when they feel like a simplified sentence was unjustly rejected, and so by

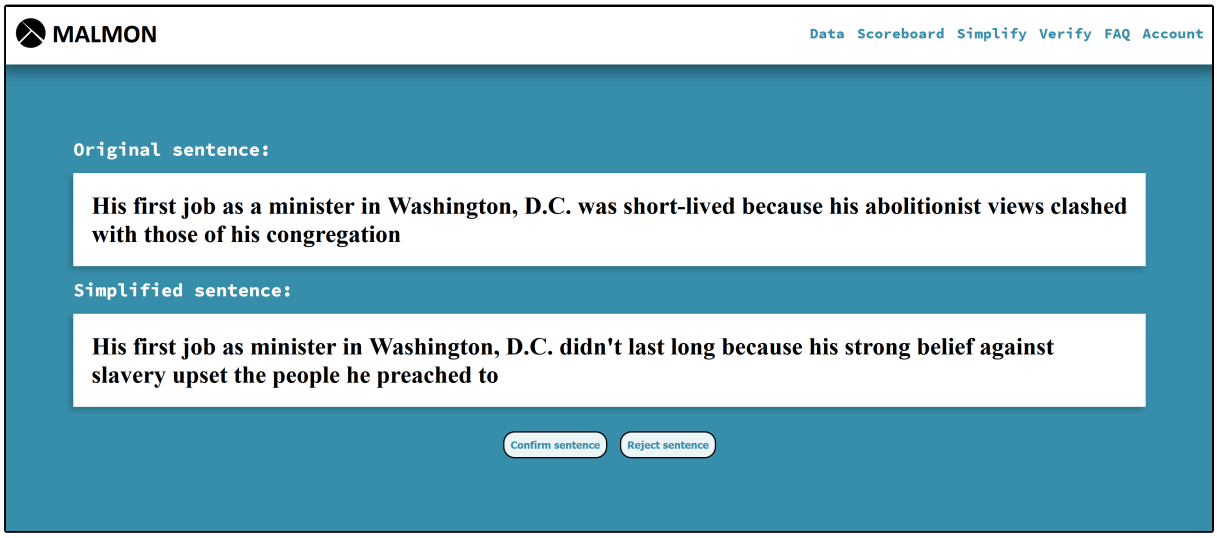

Figure 2: The verification page of the platform. Users are presented with a complex and simple sentence pair and are asked to verify the quality of the simplified sentence.

deleting the rejection it re-enters circulation and awaits confirmation by a user.

In the *user* area, an admin can view a list of all registered users on the website, again 10 at a time. The username and registration date of each user is displayed, as well as the number of simplified sentences the user has submitted and the number of verifications the user has completed. For each listed user, an admin is given the option to delete that particular user.

Then, in the *upload* area, an admin can upload a CSV file containing a list of complex sentences they wish to add to the database of the platform. These sentences will then be added to the collection of complex sentences on the platform that users are presented with.

In addition to these functionalites, an admin can also access all normal user pages and interact with the page as a user would.

4. Conclusions

We present Malmon, a data collection platform intended for crowdsourcing complex-simple sentence pairs that can be used to train automatic text simplification or ATS systems. The source code for the platform is available on Github and can be freely adapted to the needs of individual researchers. We have discussed potential use of such data, particularly in aiding people with low literacy levels, whether due to reading comprehension or cognitive disabilities, second language learners, or children. ATS can also benefit the general public when used to simplify complex, domain-specific texts, such as in the medical field, or as a preprocessing step to increase the performance of other NLP systems.

The platform combines data collection and data verification and brings it all together in combination with a simple reward system. Users can freely and easily submit simplified sentences and verify sentences from other users. Contributing to the platform evenly in both submissions and verifications increases a user's score on the scoreboard and in their account. Each user additionally receives a digital pet that grows in accordance with their score. The reward system on the site can be used by itself but it can also easily be built upon or combined with other reward systems to further incentivize user participation in the crowdsourcing process. One example of this would be a lottery-based system where users can participate in the lottery by completing the evolution of their digital pet, which can only be done by participating on the site. This could for example be done by adding a one-time-use QR code adjacent to the final stage of the digital pet.

Even though it is easy to implement other ideas with the existing reward system framework, future improvements could include additional admin functionality such as a menu for choosing reward system options and combinations. This would allow anyone to choose their preferred method of crowdsourcing without interacting with the technical side of the platform. Other possible additions to the platform worth mentioning include increasing the digital pet functionality, allowing more interaction between users, and possibly expanding the platform to allow collection of other types of data.

We hope that this platform can benefit other researchers interested in ATS, particularly those working with low-resource languages.

5. Aknowledgements

This work is co-financed by the EUROCC2 project funded by the European High-Performance Computing Joint Undertaking (JU) and EU/EEA states under grant agreement No 101101903. Parts of the work have been also supported by the European Digital Innovation Hub (EDIH) of Iceland (EDIH-IS) funded in parts by the Digital Europe Programme.

6. Bibliographical References

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

Marcelo Amancio and Lucia Specia. 2014. An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130, Gothenburg, Sweden. Association for Computational Linguistics.

Mahmoud Azab, Chris Hokamp, and Rada Mihalcea. 2015. Using Word Semantics to Assist English as a Second Language Learners. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 116–120.

Rémi Cardon and Natalia Grabar. 2020. French Biomedical Text Simplification: When Small and Precise Helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716. International Committee on Computational Linguistics.

Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification. Using Game-Design Elements in Non-Gaming Contexts. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2425–2428.

Siobhan Devlin and Gary Unthank. 2006. Helping Aphasic People Process Online Information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a Parallel Finnish–Easy Finnish Dataset from News Articles. In *1st Workshop on Open Community-Driven Machine Translation*, page 21.

Núria Gala, Anaïs Tack, Ludivine Javourey Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1353–1361.

Daniel Holmer and Evelina Rennes. 2023. Constructing Pseudo-parallel Swedish Sentence Corpora for Automatic Text Simplification. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 113–123.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C Ziegler. 2022. Simplification of Literary and Scientific Texts to Improve Reading Fluency and Comprehension in Beginning Readers of French. *Applied Psycholinguistics*, 43(2):485–512.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Poorna Kushalnagar, Scott Smith, Melinda Hopper, Claire Ryan, Micah Rinkevich, and Raja Kushalnagar. 2018. Making Cancer Health Text on the Internet Easier to Read for Deaf People who Use American Sign Language. *Journal of Cancer Education*, 33:134–140.

Lucía Ormaechea and Nikos Tsourakis. 2023. Extracting Sentence Simplification Pairs from French Comparable Corpora Using a Two-Step Filtering Method. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 30–40. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. Lexical Simplification with Neural Ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40.

Atharva Phatak. 2023. *Medical Text Simplification: Bridging the Gap Between Medical Research and Public Understanding*. Ph.D. thesis, Lakehead University.

Anna Quecke and Ilaria Mariani. 2021. How to Design Taskification in Video Games. A Framework for Purposeful Game-Based Crowdsourcing. In *CEUR Workshop Proceedings*, volume 2934, pages 1–10. CEUR-WS.

Regina Stodden and Laura Kallmeyer. 2022. TS-ANNO: An annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 145–155, Dublin, Ireland. Association for Computational Linguistics.

Ciprian-Octavian Truică, Andrei-Ionuţ Stan, and Elena-Simona Apostol. 2023. SimpLex: A Lexical Text Simplification Architecture. *Neural Computing and Applications*, 35(8):6265–6280.

Hoang Van, Zheng Tang, and Mihai Surdeanu. 2021. How May I Help You? Using Neural Text Simplification to Improve Downstream NLP Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Congcong Yang, Hua Jonathan Ye, and Yuanyue Feng. 2021. Using Gamification Elements for Competitive Crowdsourcing: Exploring the Underlying Mechanism. *Behaviour & Information Technology*, 40(9):837–854.

A. Proposed Guidelines

The following guidelines, translated to English, are aimed at crowdworkers using the platform to simplify Icelandic text. The guidelines can be changed freely by researchers using the platform so that the better suit the needs of their languages. We also include examples in English for clarity purposes.

Your task is to simplify the proposed sentences in such a way that the resulting text is better suited for readers with language difficulties (such as people that have dyslexia or aphasia), L2 speakers and/or children. When simplifying the sentences, please keep the following in mind:

- The simplified sentence should only contain common, everyday vocabulary. Please avoid specialized or uncommon words as much as possible unless the sentence explicitly explains the meaning of such words. If you are not sure whether or not the word you are using is uncommon, please refer to the following website: `https://ordtidni.arnastofnun.is/`. At the bottom of the page, you will find a frequency list for words in their base form as well as for their conjugations. You can also search for a specific word using the search bar above. If the base form of a given word has a frequency below 30.000, it should probably be avoided.
- Drop unnecessary information. The simplified sentences should maintain the meaning of the original sentences but non-important information can be omitted.
  `Example:` Snæfell er hæsta staka fjall landsins, 1833 m yfir sjó. –> Snæfell er hæsta fjall Íslands. Það er 1833 metra hátt.
  - `An example in English:` Mount Everest, is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. –> The tallest mountain in the world is Mount Everest. It is located in the Himalayas.
- Avoid unnecessary verbosity.
  `Example:` Samkvæmt ráðleggingum stofnunarinnar er mælt með því að börn hreyfi sig a.m.k. 60 mínútur á dag. –> Stofnunin mælir með því að börn hreyfi sig a.m.k. 60 mínútur á dag.
  - `An example in English:` According to the guidelines of the institution, it is recommended that children exercise for at least 60 minutes per day. –> The instutution recommends that children exercise for at least 60 minutes per day.
- Simplify sentences so that they contain as few subordinate clauses as possible. If the original sentence contains such clauses, the simplified version should rather contain multiple sentences, separated by a period.
  `Example:` Hérna er fjallið *sem mér þótti svo vænt um*. –> Hérna er fjallið. Mér þótti vænt um það.
  - `An example in English:` Watching Star Wars, which has lots of special effects, is my favorite thing to do. –> I love watching Star Wars. It has lots of special effects.
- Avoid unusual word order and stylization. Simplified sentences should preferably be in the active voice and the indicative mood.
  `Example:` Gagnrýnin *sem fram hefur komið* á fullan rétt á sér. –> Gagnrýnin *sem hefur komið fram* á fullan rétt á sér.
  - `An example in English:` Across the river and through the woods go Ella and Larry. –> Ella and Larry go across the river and through the woods.