# MC-19: A Corpus of 19th Century Icelandic Texts

**Steinþór Steingrímsson** **Einar Freyr Sigurðsson** **Atli Jasonarson**
Árni Magnússon Institute for Icelandic Studies
{steinthor.steingrimsson, einar.freyr.sigurdsson,
atli.jasonarson}@arnastofnun.is

## Abstract

We present MC-19, a new Icelandic historical corpus containing texts from the period 1800–1920. We describe approaches for enhancing a corpus of historical texts, by preparing the texts so that they can be processed using state-of-the-art NLP tools. We train encoder-decoder models to reduce the number of OCR errors while leaving other orthographical variation be. We generate a separate modern spelling layer by normalizing the spelling to comply with modern spelling rules, using a statistical modernization ruleset as well as a dictionary of the most common words. This allows for the texts to be PoS-tagged and lemmatized using available tools, facilitating usage of the corpus for researchers and language technologists. The published version of the corpus contains over 270 million tokens.

## 1 Introduction

For most areas of language technology, large text corpora and other textual resources have become increasingly important in recent years, not least due to large language models (LLMs) becoming ever more pervasive. Textual resources are not only necessary to train such models to use and decipher language, but also for question answering, information extraction and other generative tasks. With better access to data and tools to work with linguistic data, data-oriented approaches to linguistic research and lexicography have become more common and more useful, allowing more researchers to use such approaches in their work. Most commonly, large text corpora comprise recent texts. Texts from the digital era, written to be published online, can be a good tool to study recent changes and variation in language, as well as recent events and how they are perceived as they are happening. When we want to study older language, the new methods fall short if the data is lacking. In order to facilitate linguistic research for older texts, we have compiled a new corpus, the 19th Century Megacorpus (MC-19). Such research might include diachronic linguistic studies and syntactic analysis.

The aim of the MC-19 project is to compile as large a corpus as possible, comprising texts written from 1800 to 1920. The first edition of the corpus contains texts from journals and newspapers published in this period and scanned by the National and University Library of Iceland (LBS), but we intend to extend the corpus in a later edition to also include published books. We use the OCRed texts published by LBS and develop post-processing models to find and fix OCR errors in the texts, while aiming to not change anything else. Finally, we normalize the texts using modern spelling.

The contributions of the project, presented in this paper, include:

- The corpus itself, published in TEI-format[1] and in a keyword-in-context (KWIC) search engine.[2] The published corpus contains post-processed OCRed texts and a version transcribed to modern spelling, PoS-tagged and lemmatized.
- A list of common OCR-errors when processing Icelandic texts. We manually checked a wide range of random texts on *Tímarit.is* from this period and analyzed the OCR errors. The error list, available on GitHub,[3] was used for generating synthetic training data for post-processing (see Section 4.2).

[1] http://hdl.handle.net/20.500.12537/360
[2] https://malheildir.arnastofnun.is/
[3] https://github.com/stofnun-arna-magnussonar/MC19/OCRerrors



*Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 680–687

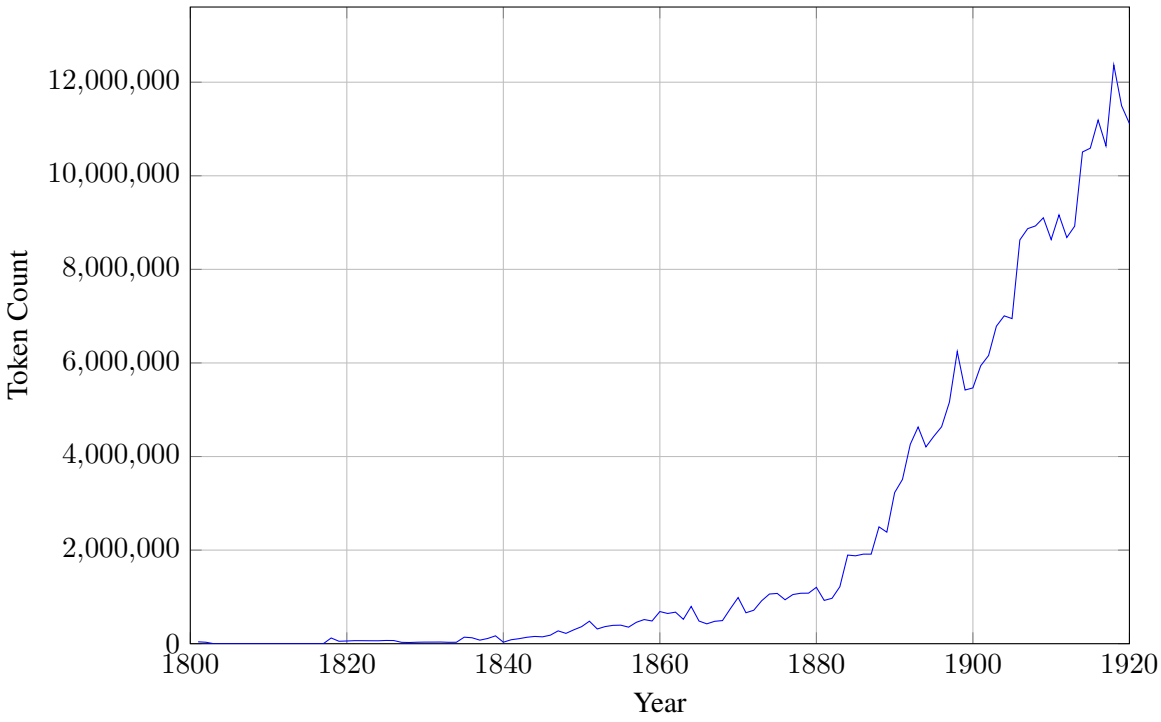


Figure 1: Token count by year (1801–1920).

- Approaches to post-processing OCR texts and transcribing to modern spelling. Models and scripts are available on GitHub.[4]

## 2 Why Do We Need a 19th Century Corpus?

Syntacticians studying Icelandic syntax, linguists studying word formation, inflectional morphology or semantics and lexicographers compiling dictionaries have been the most active users of the Icelandic Gigaword Corpus (IGC, Steingrímsson et al. 2018; Barkarson et al. 2022). Amongst these users there has been a call for corpora covering larger periods and going as far back in time as possible, in order to further the study of, for example, semantic or syntactic change. Language technologists, working on LLMs, are interested in studying how different LLMs comprehend older language in comparison with current language and to add older texts into the training process to see if it enhances the models' abilities to generate informative texts covering previous time periods. With MC-19, we aim to facilitate work in all these different fields of research.

As a demonstration of research that could be furthered with a corpus like ours, we could look at an empirical study on the reflexive passive in Icelandic conducted by Árnadóttir et al. (2011). This construction can be dated back to the 19th century as Árnadóttir et al. show. To find as old examples as they could at *Tímarit.is*, the authors had to look for word strings. To find different examples of *flýta sér* 'hurry (oneself)' in the reflexive passive, they had to search for, e.g., "var flýtt sér" ('was hurried oneself'), "var flýtt sjer", "er flýtt sér", "er flýtt sjer", "verið flýtt sér", "verið flýtt sjer", etc.; they also searched for, e.g., adverbs like *oft* 'often' intervening between the auxiliary *vera* 'be' and the participle (cf. Árnadóttir et al. 2011, 64).

This is rather time consuming, especially when one wants to look for as many different verbs as possible. This is, however, made easier in MC-19 as the corpus is PoS-tagged and lemmatized and we can therefore look for both certain word forms and tags. A search query that looks for the lemma *vera* 'be' followed by past participle (and between *vera* and the participle can be at most one word) which in turn is immediately followed by the reflexive pronoun forms *sig*/*sér*/*sín* seems to return most of the 19th-century examples from Árnadóttir et al.'s study (but of course not the ones that differ in structure from the setup in the query). This search query also returns at least two exam-

[4] https://github.com/stofnun-arna-magnussonar/MC19

ples from the 19th century that are not reported on in Árnadóttir et al. (2011).

## 3 Related Work

A wide range of historical corpora has been compiled and made available for different languages. Many of these are small, less than a million words, but there are notable exceptions. The Corpus of Late Modern English Texts (De Smet, 2005) contains over 34 million words in texts from the period 1710–1920, and the Royal Society Corpus (Kermes et al., 2016) includes all publications of the Philosophical Transactions of the Royal Society of London from 1665 to 1869, approximately 32 million tokens. ChroniclItaly (Viola, 2021) is a corpus of Italian language newspapers published in the United States between 1898 and 1920, 16.6 million words in total, and the Diorisis Ancient Greek Corpus contains 10.2 million words in texts spanning from Homer to the fifth century AD (Vatri and McGillivray, 2018). Turning to Icelandic, the Icelandic Parsed Historical Corpus (IcePaHC, Rögnvaldsson et al. 2012; Wallenberg et al. 2024) contains approximately 1 million words written between the 12th and 21st centuries. The Saga Corpus (Rögnvaldsson and Helgadóttir, 2011) contains the texts of the Icelandic sagas as well as a few other historical texts in modernized editions, and the IGC, which is 2.6 billion words in total and mostly has texts from the 21st century and the end of the 20th century, contains a few thousand words in texts written before the year 1900, all from the IGC-Law (Barkarson and Steingrímsson, 2022) subcorpus, containing law texts.

A number of studies have been carried out on how best to correct historical OCR data. Bjerring-Hansen et al. (2022) present a pipeline for correcting 19th century Danish fraktur. Their approach is rather different from ours, starting by changing "obvious and unambiguous OCR errors", then aligning multiple OCR output candidates and perform selective correction with reference to these and finally employing a spell checker.

Different approaches have been taken when doing historical spelling normalization. Schneider et al. (2017) use machine translation (MT) systems, translating original spelling into normalized texts. While they compare rule-based and SMT-based MT systems, Tang et al. (2018) evaluate the effectiveness of using neural-based MT for the task. Bollmann (2019) highlights that there is no consensus on the state-of-the-art approach to historical text normalization and compares a number of approaches. He finds that lookups based on naive memorization are most often effective for seen tokens, while MT-based methods perform best in unseen cases.

## 4 Data Processing

Our data is collected from *Tímarit.is*, a digital library platform for newspapers and periodicals that goes back to the early 19th century. The platform allows users to search texts, with OCR-generated text files for each page in the library. Rather than running our own OCR-models on the pages, which would have been resource intensive and not necessarily very beneficial, we decided to use the texts OCRed by the providers of *Tímarit.is*, LBS. In order to facilitate our work, LBS provided us with all text files for our project, covering the period in question, 1800–1920.

So that we could exclude too noisy texts, we manually checked the OCR quality of newspapers and periodicals that were candidates for our corpus. The process is described in Section 4.1.

During the selection process we compiled a list of common OCR errors. We then enlarged it by extracting a list of OCR errors from manually corrected texts from this period that we had access to. The information was used to automatically introduce OCR-like errors to correct texts, thus creating a parallel data set for training models to post-process OCRed data. We also took random samples from the texts that we decided to use and manually fixed the OCR errors to create an evaluation set. We describe this in more detail in Section 4.2.

All the selected texts were run through the post-processing models we trained, before normalizing them to modern spelling, using the approaches described in Section 4.3. Having the modern spelling variants we could PoS-tag and lemmatize the texts using the best available tools for Icelandic, which are trained on modern texts.

### 4.1 Data Selection

When selecting the publications to include, we checked all newspapers and periodicals available on *Tímarit.is* from the period 1800–1920, in total approximately 400 titles. Individual titles were evaluated by randomly selecting three volumes (years) and from each of the volumes three pages were inspected. In total, nine pages were thus

| Error | Correct | Error Word | Correct Word |
|---|---|---|---|
| 3 | ð | hú3 | húð |
| I> | Þ | I>jer | Þjer |
| ce | æ | lceknis | læknis |
| cl | d | breidcl | breidd |
| h | li | háskóh | háskóli |
| rn | m | heirnila | heimila |

Table 1: Examples of character level OCR-errors.

checked for each title. We used three categories in our evaluation:

- Green – OCR seems to be accurate and does not contain a lot of errors;
- Yellow – Most of the text looks good, but errors are common in some parts. These texts need more rigorous fixing.
- Red – Probably unusable, mostly due to OCR not giving good results. All periodicals printed in fraktur are in this category as well as texts that the OCR model fails to reproduce, commonly due to bad print or unusual layout.

In the final corpus we decided to include everything from the first two categories, green and yellow, but leave out all material in the red category, leaving us with 317 sources deemed usable.

As we performed the checks, common OCR-errors were recorded. This way, a list of 330 errors were collected, which could later be used to help with fixing the errors. Examples of this can be seen in Table 1.

### 4.2 OCR Post-processing

We carried out post-processing on all texts delivered to us by LBS, using the approaches described in Jasonarson et al. (2023). This involved using an encoder-decoder Transformer model (Vaswani et al., 2017) trained from scratch using parallel data containing OCRed texts and manual corrections of these, as well as texts populated with artificial errors in conjunction with the unspoiled data.

We had access to manually corrected texts from 19th century periodicals and journals, which we matched to the uncorrected texts.[5] This dataset

| Original | Corrected | Frequency |
|---|---|---|
| *p* | *þ* | 2,779 |
| *i* | *í* | 1,141 |
| *li* | *h* | 247 |
| *rn* | *m* | 166 |
| *m* | *rn* | 77 |

Table 2: Examples of automatically extracted errors and statistics on them.

contains in total over 2 million tokens. We also used this data to gather more examples of OCR errors and to create statistics on which errors are the most common, examples of which are shown in Table 2. In turn, this information was used to generate a new dataset containing artificial errors.

The data into which the artificial errors were inserted were texts published between 1830 and 1920, taken from the Icelandic Text Archive.[6] By doing this we have parallel data, with correct texts on the one hand and the same texts with errors like the ones commonly found in OCR output on the other. This data can then be used to train a system that effectively translates erroneous texts to correct texts, fixing many errors like the ones found in Table 1. In total, the artificial corpus contained almost 3 million tokens. We combined our two parallel datasets and split it into training and validation data, with the validation data being 15% of the total, approximately 750 thousand tokens, and the training set approximately 4.2 million tokens.

To evaluate the post-processing accuracy, we created an evaluation set by selecting random pages from the corpus and manually correct them. The evaluation set contains in total 18k tokens.

We trained three models, as described in Jasonarson et al. (2023), the best being a fine-tuned version of ByT5-base (Xue et al., 2022) which achieved a word error rate reduction of 55.07% – cutting the number of erroneous words in half.

### 4.3 Modernizing the Spelling

We manually modernized the 10,000 most common words in our training data and created a lookup dictionary. We also built a statistical spelling modernization ruleset by iterating over a small, manually modernized, parallel corpus, one token at a time, extracting the necessary ed-

[5]These texts are a product of the project *Language Change and Linguistic Variation in 19th-Century Icelandic and the Emergence of a National Standard*, led by Ásta Svavarsdóttir at the Árni Magnússon Institute for Icelandic Studies (e.g. Svavarsdóttir et al. 2014).

[6]https://clarin.is/en/resources/textarchive/

its needed to convert an old token into a modern one. This resulted in 101 rules, such as *je→é* and *p→f*, both of which are a frequent change from old tokens to modern ones.

To modernize our corpus, our system iterates over every sentence in a given original text and generates a modern counterpart. It looks at every token in the original sentence and checks whether it exists in the Database of Icelandic Morphology (DIM, Bjarnadóttir et al. 2019). If it does, the token gets added unchanged to the new modern sentence. If the token is not found in DIM, the system checks whether the word exists in the manually corrected lookup dictionary, and if so, the modernized spelling variant gets added to the new sentence. If a token is shorter than 3 characters, we do not try to modernize it and simply add it to the new sentence.

If an original token's modern counterpart has not been found at this point, we create an empty list, which we populate with plausible candidates that we produce with several methods.

1. Using Kvistur (Daðason et al., 2020), we check whether the token is a compound word. If all of its parts exist in DIM, we add it to the candidate list.

2. We check whether there is a word in DIM that has a Levenshtein-distance (Levenshtein, 1966) of 1 (or 2, if the token is 12 characters or longer) from the original token. If it does and its edit from the original token is found in our statistical spelling modernization ruleset, we add it to the list, e.g. if the original token is *eptirlegukind* and *eftirlegukind* is found in DIM, as *p→f* is a known spelling modernization rule.

3. We apply all of the possible modernization rules to the token and if any of them produces a token which exists in DIM, we add it to the list.

4. We edit the token with two rules. If it ends with 'r', we try adding 'u' in front of it, e.g. *hestr→hestur*, and check whether the resulting token is found in DIM. (Older forms of nouns often do not have 'u' in the ending before 'r'.) We also check if doubling a consonant in the token, e.g. *bygð→ byggð*, results in a known modern token. If either of these returns a known modern token, we add it to the list of plausible candidates.

5. We use two models, a modern GEC model[7] and IceBERT.[8] We use the former as a spellchecker to edit the current token, and the latter, by masking the current token, to guess which token should be in its place. If either of these returns a token, which, when compared to the original token, can be inferred from the rules in our statistical spelling modernization dataset, we add it to the candidate list.

When all of these checks are completed, we simply add the most suggested token to the new sentence. If all of these methods fail, however, in producing a plausible candidate, the original token stays in the modern sentence. In such a case the token could be an uncommon one, but free of errors, or it could be the case that the applied methods fail to suggest the correct form.

### 4.4 Tagging and Lemmatization

The most accurate PoS-tagger and lemmatizer for Icelandic are trained to work with modern spelling varieties. We thus only tag and lemmatize the normalized version of the texts. We start by tokenizing the texts using Tokenizer,[9] a Python program developed for tokenizing Icelandic texts. We use ABLTagger 3.0.0 (Steingrímsson et al., 2019; Jónsson et al., 2021) for PoS-tagging the texts. The tagger is reported to have an accuracy of 96.7% when using cross-validation on MIM-GOLD (Helgadóttir et al., 2014; Barkarson et al., 2021), the standard dataset for training and evaluation of PoS-tagging for Icelandic. Nefnir (Ingólfsdóttir et al., 2019) is the most suitable lemmatizer for Icelandic texts, reported to produce only a fraction of the errors other lemmatizers for Icelandic produce. It uses the tags output by the PoS-tagger to help with finding correct lemmas, using suffix substitution rules derived from DIM.

### 4.5 Data Statistics

MC-19 contains a total of 272,516,487 tokens from 317 sources. As shown in Figure 1, most of the tokens are from material published late in the

[7] ByT5-model: https://huggingface.co/mideind/yfirlestur-icelandic-correction-byt5

[8] https://huggingface.co/mideind/IceBERT

[9] https://pypi.org/project/tokenizer/

| Title | Period | Token Count |
|---|---|---|
| Lögberg | 1888–1920 | 41,002,958 |
| Heimskringla | 1886–1920 | 32,486,522 |
| Þjóðólfur | 1848–1920 | 15,364,734 |
| Morgunblaðið | 1913–1920 | 13,175,574 |
| Lögrétta | 1906–1920 | 8,485,516 |
| Austri | 1883–1888; 1891–1917 | 7,183,953 |
| Fjallkonan | 1884–1911 | 6,993,309 |
| Þjóðviljinn + Þjóðviljinn ungi | 1886-1915 | 6,971,801 |
| Skírnir | 1827–1920 | 6,460,688 |
| Norðurland | 1901–1920 | 5,105,768 |

Table 3: The ten publications in MC-19 that contain the largest number of tokens. The table shows the period as represented in the corpus. Some of these publications continued to be published after 1920.

period, with more than 50% being from the last 14 years (1907–1920). The first 50 years only contain approximately 3.5 million tokens (there is no data in the corpus for the years from 1803 to 1817).

Furthermore, a few publications tower over the rest, with ten publications containing more than 5 million tokens each, as shown in Table 3. These ten publications represent more than half the corpus data.

## 5 Use and Availability

The corpus is published under an open CC BY 4.0 license. It is available online in two different forms for different uses and users. It is made available for search online in a KWIC-portal, powered by KORP (Borin et al., 2012). Users can search for word forms in both the original version (OCRed text) and in the modern spelling transcription, with the modern spelling transcription being PoS-tagged and lemmatized, allowing for more complicated search in that data. The results are shown in parallel, so while the user can search using modern spelling varieties, the original ones are also shown. This format is expected to mostly be useful to linguists, lexicographers and students of Icelandic.

The TEI-version is available for download. It contains whole sentences in the original version as well as the normalized version using modern spelling. The normalized version is furthermore PoS-tagged and lemmatized. We expect this format to be most useful for language technologists for analyzing and building tools and language models. Linguists competent in programming may also find that working with these annotated documents allows for more complicated analysis and research than when limited to KWIC-analysis.

## 6 Conclusion and Future Work

We have presented a new text corpus, MC-19, containing Icelandic texts from the 19th century and the first decades of the 20th century. The first version of this corpus has been published and is made available in a TEI-format as well as in an online KWIC-platform, powered by Korp.

While care has been taken to make the texts as readable and close to the printed material as possible, using a post-processing step and a spelling-modernization step, there is always room for improvement. The post-processing process reduces the number of OCR errors by 55.07%. Improving the performance in this step would make the corpus more accurate and useful. This could possibly be achieved by improving the post-processing models, for example by generating more artificial training data or more diverse training data. Some error reduction may be achieved simply by replacing possible errors with possible corrections, using our error list. For such an approach, which tends to be greedy, some measures would need to be taken to limit the possibility of generating new errors. This could possibly be achieved by mapping only from unknown words (containing possible errors) to known words, calculating the likelihood of the change using n-grams or perplexity calculations or other approaches that may prove useful.

While most of the sentences in the corpus are as printed in the original publications, some are garbled due to problems with OCR that our methods could not solve. Training a classifier to select bad sentences for removal could make the corpus an

even better tool.

The spelling-modernization step helps the user find common words for which the spelling has changed, allowing for easier search and usage of the corpus, but the user will still find that some words are not modernized. A more thorough examination of this and improvements in the process will help with using the corpus for research. We intend to revisit these steps for a future version of the corpus, integrating additional normalization techniques and manually evaluate the merits of different approaches to this problem. We also intend to add texts from books published in the period, and are working on OCR-reading fraktur texts. While these texts may not add very much to this corpus in terms of word count, as the bulk of published texts in the period is in newspapers and periodicals, it may show a greater variety, both in terms of language and content. Available texts from previous periods, printed and hand-written, are also being considered for a sister corpus to this one.

## Acknowledgments

We would like to thank three anonymous reviewers for helpful comments. We would like to thank the following people for their collaboration and contributions to the project: Finnur Ágúst Ingimundarson and Árni Davíð Magnússon for analyzing the OCR errors and Starkaður Barkarson for his work on publishing the corpus in TEI-format and setting it up on Korp. This project was supported by Rannís Infrastructure Fund, grant number 200336-6101.

## References

Hlíf Árnadóttir, Thórhallur Eythórsson, and Einar Freyr Sigurðsson. 2011. The passive of reflexive verbs in Icelandic. *Nordlyd*, 37:39–97.

Starkaður Barkarson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Árni Davíð Magnússon, Kristján Rúnarsson, Steinþór Steingrímsson, Haukur Páll Jónsson, Hrafn Loftsson, Einar Freyr Sigurðsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir. 2021. MIM-GOLD 21.05. CLARIN-IS.

Starkaður Barkarson and Steinþór Steingrímsson. 2022. IGC-Law 22.10 (annotated version). CLARIN-IS.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. Evolving large text corpora: Four versions of the Icelandic Gigaword Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts. a heuristic procedure for correcting OCR data. *Digital Humanities in the Nordic and Baltic Countries Publications*, 4(1):177–186.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey. European Language Resources Association.

Jón Daðason, David Mollberg, Hrafn Loftsson, and Kristín Bjarnadóttir. 2020. Kvistur 2.0: a BiLSTM compound splitter for Icelandic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3991–3995, Marseille, France. European Language Resources Association.

Hendrik De Smet. 2005. A corpus of Late Modern English texts. *ICAME Journal*, 29(2005):69–82.

Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2014. Correcting errors in a new gold standard for tagging Icelandic text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2944–2948, Reykjavik, Iceland. European Language Resources Association.

Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland. Linköping University Electronic Press.

Atli Jasonarson, Steinþór Steingrímsson, Einar Freyr Sigurðsson, Árni Davíð Magnússon, and Finnur Ágúst Ingimundarson. 2023. Generating errors: OCR post-processing for Icelandic.

In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 286–291, Tórshavn, Faroe Islands. University of Tartu Library.

Haukur Páll Jónsson, Hrafn Loftson, and Steinþór Steingrímsson. 2021. ABLTagger (PoS) - 3.0.0. CLARIN-IS.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey. European Language Resources Association.

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2011. Morphological tagging of Old Icelandic texts and its use in studying syntactic variation and change. In *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 63–76, Berlin/Heidelberg. Springer.

Gerold Schneider, Eva Pettersson, and Michael Percillier. 2017. Comparing rule-based and SMT-based spelling normalisation for English historical texts. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 40–46, Gothenburg. Linköping University Electronic Press.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria. INCOMA Ltd.

Ásta Svavarsdóttir, Sigrún Helgadóttir, and Guðrún Kvaran. 2014. Language resources for early Modern Icelandic. In *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage*, pages 19–25, Reykjavik, Iceland.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5999–6009, Long Beach, California.

A. Vatri and B. McGillivray. 2018. The Diorisis Ancient Greek Corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55 – 65.

Lorella Viola. 2021. ChroniclItaly and ChroniclItaly 2.0: Digital heritage to access narratives of migration. *International Journal of Humanities and Arts Computing*, 15(1–2).

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2024. IcePaHC 2024.03 – A significant treebank upgrade. In *CLARIN Annual Conference Proceedings*, pages 168–171, Barcelona, Spain. ISSN 2773-2177 (online).

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.