Creating an Error Corpus: Annotation and Applicability

Þórunn Arnardóttir
University of Iceland
Reykjavík, Iceland
thar@hi.is

Xindan Xu
University of Iceland
Reykjavík, Iceland
xindanxu@hi.is

Dagbjört Guðmundsdóttir
University of Iceland
Reykjavík, Iceland
dagu@hi.is

Lilja Björk Stefánsdóttir
University of Iceland
Reykjavík, Iceland
lbs@hi.is

Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

In this paper, we describe the Icelandic Error Corpus, a manually annotated error corpus for Icelandic. The Icelandic Error Corpus consists of texts from three sources: student essays, online news and Wikipedia articles, with a total of 56,794 annotated error instances. The corpus is used to analyze errors made by Icelandic native speakers, which are in turn used to guide the development of an Icelandic open-source spellchecker. The corpus is delivered in an augmented TEI format and published under an open-source license.

1 Introduction

The Icelandic Error Corpus is a collection of texts in modern Icelandic which are manually annotated for errors related to spelling, grammar, and other issues. The corpus consists of three genres: student essays, online news and Wikipedia articles. In total, the corpus consists of 4,044 texts with 44,268 revision spans and 56,794 categorized error instances. It is published under a CC BY 4 license and is available from the Icelandic CLARIN repository (Ingason et al., 2021).

A manually annotated error corpus is a useful resource for various tasks within language technology. It can be used to analyze real-world spelling and grammar errors, which in turn can be used to guide the development of a spellchecker. The Icelandic Error Corpus was created for this purpose and it is a novel kind of resource in the context of Icelandic. It reflects the mistakes that Icelandic informants make in written text and is used to measure and improve the performance of an automatic spelling and grammar corrector for Icelandic.

The paper is structured as follows. Section 2 discusses error corpora in general and currently available Icelandic spellcheckers. Section 3 describes the text sources used for the corpus while Section 4 describes the annotation process and Section 5 the annotation scheme. Section 6 gives an overview of the Icelandic Error Corpus and reports on statistical information on it, and we then conclude with Section 7.

2 Error Corpora and Icelandic Spellcheckers

Spelling and grammatical error correction are established tasks within natural language processing. Different methods are available for doing so, some of which are based on an error corpus, a collection of texts which have been annotated for errors. Error corpora can be generated automatically by comparing the edit history of texts (Grundkiewicz and Junczys-Dowmunt, 2014) or by identifying typo edits using a trained classifier (Hagiwara and Mita, 2020). They can also be created by manually annotating text,

which is the case for the Birbeck spelling error corpus (Mitton, 1980). The data comprising an error corpus also differ, some of them consisting of texts written by native speakers (Deksne and Skadina, 2014; Rosner et al., 2012) and others consisting of texts written by non-native speakers (Boyd et al., 2014; Tenfjord et al., 2006; Volodina et al., 2016). Errors within a corpus depend on the data it consists of, which can either be texts written by informants or word lists. An important task for a spellchecker is context-sensitive correction, especially when strings are pronounced the same but are semantically distinct, as in the English pair *there/their*. An Icelandic corpus of such strings already exists (Friðriksdóttir and Ingason, 2020b) and can be used along with a general error corpus, made up of real-world texts, when developing an Icelandic spellchecker. This corpus has been used in recent experiments involving different types of binary classifiers (Friðriksdóttir and Ingason, 2020a; Friðriksdóttir and Ingason, 2020b), expanding on earlier research that depended on more limited data sets (Ingason et al., 2009).

A few Icelandic spellcheckers exist, but they differ with respect to their accessibility as well as the set of features they implement. The Skrambi system is available through an online user interface[1] and is capable of context-sensitive spellchecking (Daðason, 2012). Another spellchecker, Púki, is only available through a fee. It includes a thesaurus and can therefore suggest synonyms for words in a text and learn new words and terms from the text itself. The only open-source Icelandic spellchecker is GreynirCorrect,[2] which is published under the MIT license. The tool returns both errors and suggestions on spelling and grammar. The Icelandic Error Corpus is used to improve the performance of this spellchecker by analyzing real-world examples of spelling and grammatical errors.

3 Data

Three text genres were used in the corpus: student essays, online news texts and Wikipedia articles. These sources were chosen for two main reasons: first, they reflect different styles of writing; and second, they are readily available because they have already been compiled and published, without annotation, as part of the Icelandic Gigaword Corpus (Steingrímsson et al., 2018). The student essays were written by high school students between the age of 16 and 20. These texts were published anonymously in the Icelandic Gigaword Corpus under a license that imposes certain restrictions on derived resources. Therefore, sentences within these texts had to be shuffled before they could be released under an open-source license. Texts in the online news and Wikipedia articles were published under an open-source license in the Icelandic Gigaword Corpus and therefore, they did not need to be shuffled before they could be published as part of the error corpus.

As mentioned in Section 2, the Icelandic Error Corpus is used for developing GreynirCorrect. It was therefore split into a training corpus and a test corpus, which allows the developers to measure the spellchecker's performance on data which the developers have not seen. Random sampling was used to split the corpus into a development corpus, 90% of the total, and a test corpus, the other 10%. Section 6 reports on the number of files and errors in the respective parts of the error corpus.

4 Annotation Process

The annotation process uses a layered approach which culminates in a collection of augmented TEI-format XML documents with the eventual error annotations. The process consists of five steps: text cleanup, proofreading, conversion to TEI-format XML, error code labeling and format checks.

First, a text is converted to the appropriate format, i.e. the XML-format files of the Icelandic Gigaword Corpus are converted to text format, and any extra information is removed. The second step in the process involves manual proofreading and correction using any tool that allows for correcting errors and preserving the original version of the text. Microsoft Word and its Track Changes feature were used for this purpose. The annotators working within this step, 8 in total, were solely proofreaders and could therefore specialize in this task, allowing for more precise and consistent corrections.

After the texts have been proofread, the incorrect and correct versions of each document are aligned and merged. This is done using a Python script which results in an XML structure that explicitly marks

[1] `http://skrambi.arnastofnun.is`
[2] `https://github.com/mideind/GreynirCorrect`

every correction made to the text, by using a revision span. The next step in the process consists of manual annotation of the errors, whereby annotators work with the XML structure, labeling each error with one or more error codes. The annotators working on this step were separate to the proofreaders and specialized in error code labeling. The final step, before publication, is checking each file's format, i.e. ensuring that the XML format is readable and that all labeled error codes are part of the annotation scheme.

5 Annotation Scheme

The annotation scheme developed for the corpus consists of three hierarchical levels: main categories, subcategories, and error codes used during annotation. The annotation scheme is similar to that used in the MERLIN corpus (Boyd et al., 2014). The main categories are six in total, the subcategories are 31 and the error codes are 253.[3] The annotation scheme evolved as more texts were annotated, being descriptive in that it reflects the errors which appear in the corpus and none beyond that. The error codes, the lowest level of the annotation scheme, are precise and there is a clear correspondence between an error and an error code, while the subcategories, the middle level, better reflect the error types in general, e.g. agreement errors, typographical errors, etc.

A group of four annotators, separate to the ones who proofread the texts, worked on error code labeling and created the annotation scheme simultaneously. We believe that this separation between proofreaders and annotators ensures more precise corrections, and it is in contrast to the approach taken in Deksne and Skadina (2014) and Rosner et al. (2012), where proofreaders also annotated the errors. Furthermore, the error annotation in Deksne and Skadina focuses solely on spelling errors and foreign words while the annotation scheme in Rosner et al. is similar to the one used in the Icelandic Error Corpus, only simpler. The first texts in the corpus were annotated by all annotators and then reviewed to ensure that the annotation was agreed upon. Additionally, all annotators had to agree on adding a new error code to the scheme.

Three steps were taken to revise the annotation scheme. First, specialists in language use consultation and spellchecking were consulted. As a result, error codes were refined and redundant error codes were merged or removed from the annotation scheme. Second, 10% of all instances of each error code was sampled and reviewed by the annotators. If a particular error code was incorrectly used for more than 33% of the cases in the sample, all instances of the error code were manually reviewed and corrections made. Third, all instances of each error code are reviewed while developing the spellchecker and corrections made when necessary. All steps lead to both a more refined annotation scheme and more accurate error code labeling.

6 Overview of the Error Corpus

The Icelandic Error Corpus consists of 4,044 texts, which were processed and annotated for errors. A total of 44,268 revisions were made and 56,794 errors annotated. These two numbers are different because a revision span can include more than one error. Table 1 shows the number of files, revisions and categorized error instances in each subcorpus and their respective text genres in the Icelandic Error Corpus. The corpus is delivered in augmented TEI-format XML documents, and is therefore machine-readable. As a result, some corpus management platforms particular to TEI-format files can be used to obtain information from the corpus.

The overall average number of errors per 1000 words in the Icelandic Error Corpus is 45.76. However, there are clear differences in the error rates between genres within the corpus. As is shown in Table 1, the number of errors per 1000 words is lowest in the online news, and highest in the Wikipedia articles. In the development corpus, the number of errors per 1000 words is similar between the online news (35.74) and student essays (37.83), whereas the number of errors per 1000 words is substantially higher in the Wikipedia articles (62.03; Table 1). This trend is also seen in the test corpus, although the number of errors per 1000 words is slightly higher for student essays, and slightly lower for Wikipedia articles.

[3]The complete annotation scheme is available at `https://github.com/antonkarl/iceErrorCorpus/blob/master/errorCodes.tsv`

Subcorpus	Files	Revisions	Categorized Errors	Errors/1000w
Development corpus				
Student essays	158	4,719	5,947	37.83
Online news	2,638	15,969	19,579	35.74
Wikipedia articles	881	20,216	26,786	62.03
Test corpus				
Student essays	18	645	828	43.30
Online news	267	1,334	1,663	32.74
Wikipedia articles	82	1,385	1,991	58.03
Total	4,044	44,268	56,794	45.76

Table 1. Overview of the number of files, revision spans and categorized error instances in both parts of the Icelandic Error Corpus.

Table 2 shows the 10 most common subcategories in the Icelandic Error Corpus, as indicated by the first column. It also lists the most common error codes within each subcategory, ordered by frequency, the subcategory's frequency and its proportion of all errors in the corpus. The most common error type is incorrect use of punctuation, such as when wrong quotes are used. This amounts to 25% of all the errors in the corpus. The second most prominent error type is "wording", which comprises 15% of all errors in the corpus. The remaining subcategories shown in Table 2 have a substantially lower frequency, with a proportion ranging from 7% to 3%.

Subcategory	Main category	Most common error codes	Freq	Prop (%)
punctuation	orthography	wrong-quots, extra-comma, missing-comma	13,357	25.46
wording	style	wording	7,734	14.74
spacing	orthography	missing-hyphen, split-compound, merged-words	3,663	6.98
nonword	orthography	nonword, compound-collocation	3,203	6.11
typo	orthography	missing-letter, letter-rep, extra-letter	2,981	5.68
style	style	nonit4it, it4nonit, fw4ice	2,920	5.57
insertion	vocabulary	extra-word, extra-words	2,885	5.50
syntax	grammar	missing-fin-verb, missing-sub, missing-obj	2,244	4.28
omission	vocabulary	missing-word, missing-words	1,763	3.36
capitalization	orthography	upper4lower-common, lower4upper-proper, upper4lower-proper	1,695	3.23

Table 2. Most common error types in the Icelandic Error Corpus.

7 Conclusion

In this paper, we have described the Icelandic Error Corpus, an open-source collection of texts which have been annotated for errors, and its purpose in developing an Icelandic spellchecker. The corpus consists of three text genres: student essays, online news and Wikipedia articles, which have been annotated for errors regarding spelling, grammar and other issues. The error corpus is published in an augmented TEI format, with revision spans marking the corrections made to a text and error codes for categorizing each error. In total, the corpus consists of 44,268 revision spans and 56,794 categorized error instances.

This manually annotated error corpus is important, not only for developing an open-source spellchecker, but also to depict real-world spelling and grammar errors which Icelandic informants make. The corpus facilitates the development of a spellchecker that takes into account the needs of native Icelandic speakers, so that it can detect and correct errors which are often produced by them.

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.

References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., and Vettori, C. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavík, Iceland, May. European Language Resources Association (ELRA).

Daðason, J. 2012. Post-correction of Icelandic OCR text.

Deksne, D. and Skadina, I. 2014. Error-annotated corpus of Latvian. In *The Sixth International Conference "Human Language Technologies – The Baltic Perspective" (Baltic HLT 2014)*, pages 163–166, 09.

Friðriksdóttir, S. R. and Ingason, A. K. 2020a. Disambiguating confusion sets as an aid for dyslexic spelling. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 1–5, Marseille, France, May. European Language Resources Association.

Friðriksdóttir, S. R. and Ingason, A. K. 2020b. Disambiguating confusion sets in a language with rich morphology. In *Proceedings of ICAART 12 (International Conference on Agents and Artificial Intelligence)*.

Grundkiewicz, R. and Junczys-Dowmunt, M. 2014. The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In Przepiórkowski, A. and Ogrodniczuk, M., editors, *Advances in Natural Language Processing*, pages 478–490, Cham. Springer International Publishing.

Hagiwara, M. and Mita, M. 2020. GitHub typo corpus: A large-scale multilingual dataset of misspellings and grammatical errors. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6761–6768, Marseille, France, May. European Language Resources Association.

Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. 2009. Context-sensitive spelling correction and rich morphology. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 231–234, Odense, Denmark, May. Northern European Association for Language Technology (NEALT).

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., and Xu, X. 2021. Icelandic error corpus (IceEC) version 1.1. CLARIN-IS.

Mitton, R. 1980. Birkbeck spelling error corpus. Oxford Text Archive.

Rosner, M., Gatt, A., Attard, A., and Joachimsen, J. 2012. Incorporating an error corpus into a spellchecker for Maltese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 743–750, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Tenfjord, K., Meurer, P., and Hofland, K. 2006. The ASK corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G., and Sandell, M. 2016. Swell on the rise: Swedish learner language corpus for European reference level studies. *CoRR*, abs/1604.06583.