CLARIN Annual Conference 2021

PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic

Isidora Gliśić
University of Iceland
Reykjavík, Iceland
isgl4@hi.is

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

Abstract

The Icelandic L2 Error Corpus is an expanding collection of texts written by users of Icelandic as a second language, published on CLARIN. It currently consists of 17508 manually-annotated errors in different categories pertaining to grammar, spelling, lexical and other issues. The corpus was used to perform a contrastive interlanguage analysis using a native speaker reference corpus comparing it to the Icelandic Error Corpus. This paper presents the corpus and the first results of the analysis.

1 Introduction

The popularity of Icelandic as a learner language is a quite novel phenomenon and teaching materials are still in development. With new language technology efforts in Iceland, it is finally possible to create ICALL (Intelligent Computer-Assisted Language Learning) solutions for Icelandic and a major step towards this is creating a learner error corpus. At the moment of writing, the Icelandic L2 Error Corpus is a collection of 70 texts, predominantly student essays, annotated for various types of errors. The corpus contains a total of 12081 revision spans and 17508 error instances, where a revision span is a word or a span of words that have been corrected in the annotation process and an error instance is a link between a revision span and a categorization of an error found in the span. This corpus is still growing and will be utilized in analysing learners’ interlanguage for the purpose of perfecting teaching materials (both electronic, textbooks and syllabi) and automatic correction tools.

The paper is structured as follows. Overview of previous research on learner interlanguage for Icelandic and the introduction to the new Icelandic L2 corpus is in section 2. Section 3 describes the methods that we used. Section 4 presents the results of a comparative analysis with the Icelandic general error corpus.

2 Resources for studying L2 errors in Icelandic

Until relatively recently, not many foreigners were interested in learning Icelandic and no textbooks or teaching methodology existed. The first contrastive analysis of the learner language emerged in the 1980s (Sigmundsson, 1987), and it was not until the ‘90s that the first textbooks started being published and even more recently that attention was drawn to learner errors (Porvaldsdóttir and Garðarsdóttir, 2013; Ólafsson, 2016). Finally the learner corpus has been published (Ingason et al., 2021) and is at this moment still in development. The corpus in its current form is provided in open access via the CLARIN repository and GitHub.

The Icelandic L2 Error Corpus currently consists of 70 texts from 27 adult second language speakers of Icelandic (mostly aged between 20–40) with 13 different first languages, containing 17508 categorized error instances. Further analysis of the corpus data will follow in section 4. The texts are previously unpublished and obtained directly from their authors, who choose whether the text is to be published

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

1The corpus is available at: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/106

Proceedings CLARIN Annual Conference 2021
under their name or anonymously. The texts are for the most part student essays submitted for evaluation in various courses at the university. The mean number of words per text is 1780 but this number varies when separated by skill level (with the mean for A1 texts being 324 words and 5177 words for C2) as the nature and type of texts vary – the highest skill level texts typically being long academic essays and parts of or entire MA theses. More numerical data based on skill level is depicted in Table 1.

<table>
<thead>
<tr>
<th>Level</th>
<th>Files</th>
<th>Total words</th>
<th>Total errors</th>
<th>Errors/1000w</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>12</td>
<td>3889</td>
<td>1095</td>
<td>281.56</td>
</tr>
<tr>
<td>A2</td>
<td>19</td>
<td>11901</td>
<td>2532</td>
<td>212.76</td>
</tr>
<tr>
<td>B1</td>
<td>8</td>
<td>10439</td>
<td>1919</td>
<td>183.83</td>
</tr>
<tr>
<td>B2</td>
<td>11</td>
<td>19504</td>
<td>3367</td>
<td>172.63</td>
</tr>
<tr>
<td>C1</td>
<td>9</td>
<td>21940</td>
<td>2715</td>
<td>123.75</td>
</tr>
<tr>
<td>C2</td>
<td>11</td>
<td>56953</td>
<td>2911</td>
<td>103.79</td>
</tr>
</tbody>
</table>

Table 1: Total number of files, words, errors and errors per 1000 words per skill level.

The advantage of using student essays is the accessibility of texts from subjects with different first and second language background. Furthermore, it is also relatively easy to estimate their proficiency level based on how far along they are in their studies. On the other hand, due to the nature of the texts (academic essays) some types of errors tend to be more prevalent than in other types of writing. Apart from that, many generic errors may have been corrected already as the texts tend to be polished for better academic success. Additionally, as participation is strictly voluntary, it has been challenging to obtain sufficient amount of texts to be able to draw any conclusions based on their demographic features.

The standard for assessing the stage of learners’ interlanguage is the Common European Framework of Reference for Languages (CEFR) (Piccardo et al., 2018). The scale is particularly important in evaluating learner errors, as specific types of errors typically emerge on specific proficiency levels, with typical stagnation and regression points (Thewissen, 2013). The Icelandic as a second language program is separated into a one-year Practical diploma in Icelandic which covers the proficiency level A1–A2 and a 3-year bachelor degree where the students are estimated to be on the level B1–B2 by the end of the first year, and reach B2—C1 by the end of the program (Garðarsdóttir and Porvaldsdóttir, 2020).

How the corpus was built and the process of extracting and analyzing relevant data will be explained in the next section.

3 Methods

The texts for the Icelandic L2 Error Corpus were collected through an online consent form and manually proofread and annotated for errors. Microsoft Word’s tracking changes feature was used for this because it preserves both the original version and the corrected version. Both versions of the text were extracted and converted, using a Python script, into a single augmented TEI format XML document with labeled enumerated sentences, words and punctuation, and revision spans with unique id numbers containing errors. The errors were analysed and annotated manually and the annotators would label one or several error codes in each revision span. Figure 1 shows an example of a revision span containing several error codes. It demonstrates that a revision span can have both multiple codes for different errors, as well as codes which apply to the same error in which case they share the same index (idx). So in this example, the errors with the id "255-1" refer to the first two words in the revision span, whereas the last word needs to be covered by a different error type.

The annotation system used for error labeling was originally developed for the Icelandic Error Corpus (Ingason et al., 2020) previously released through CLARIN, which contains errors in native speaker texts,
and later expanded with new labels that were specific to the L2 errors. The error tagset consists of 19 categories which are further divided into subclasses. Some subclasses are very narrow while others are more wide-ranging (notably the Grammar and Punctuation category) and in total there are 259 error codes. A list of all the codes along with an example and a description is available at https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/IEC_ErrorCodes.pdf.

After the dataset of TEI documents has been finalized, statistical analyses are conducted that include quantifying the number of texts, revision spans and error occurrences in the corpus, as well as contrasting the L2 error corpus with the Icelandic Error Corpus by ranking the frequency of the error codes extracted as the number of errors per 1000 words. Moreover, each document contains metadata including the author’s first language, other languages, length of residence in Iceland, length of study of Icelandic, and proficiency level. This data is stored to extract specific information on errors based on these parameters that can be used in further research.

4 Data analysis

The method we used is contrastive interlanguage analysis (CIA) which compares varieties within one language using two types of comparison: comparing learner language with native speaker reference corpora (L2 vs. L1) or comparing different varieties of learner language (L2 vs. L2) (Granger, 2008). The former can uncover the distinguishing features of L2 language use while the latter allows us to assess the generalizability of interlanguage features across different factors, learner and task based. As an error corpus for L1 Icelandic has recently been published (Ingason et al., 2020), this provides us with the possibility to make a CIA based on the first type mentioned.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Files</th>
<th>Total words</th>
<th>Revisions</th>
<th>Categorized Errors</th>
<th>Errors/1000w</th>
</tr>
</thead>
<tbody>
<tr>
<td>Icelandic Error Corpus</td>
<td>4046</td>
<td>1137941</td>
<td>44261</td>
<td>55346</td>
<td>44.56</td>
</tr>
<tr>
<td>Icelandic L2 Error Corpus</td>
<td>70</td>
<td>124626</td>
<td>12081</td>
<td>17508</td>
<td>140.73</td>
</tr>
</tbody>
</table>

Table 2: Numerical data for both L1 and L2 Icelandic error corpora.

As Table 2 demonstrates, the number of errors per 1000 words is significantly higher in L2 texts than in the general corpus. This is not surprising as learner errors are quite frequent, and particularly on lower proficiency levels the text can be so convoluted and inaccurate that making revisions proved to be a challenge as sometimes entire sentences needed to be rewritten for the text to be semantically coherent. However, it must be noted that the learner error corpus contains significantly fewer and less genre-diverse texts and this may affect how our findings generalize to L2 users as a population.
The analysis also sheds light on a significant disparity in the frequency of certain error categories and subcategories in L2 Icelandic compared to L1 errors. The most frequent error categories in the L2 corpus are: grammar (43.57%), punctuation (12.14%) and wording (11.63%). Each other error category comprises 5% or less of total errors. In comparison, the category grammar accounts for only 11.8% in the general Icelandic Error Corpus. There are 35 error codes that appear only in the L2 corpus, 27 of which are unsurprisingly within the grammar category.

These errors mostly involve case government (case-verb, case-collocation, case-prep, case-adj) as it is not intuitive in the language learning process which case is governed by a certain preposition or verb. For example, L2 users will commonly misuse a phrasal verb and instead of interpreting it as a verb+particle combination they would take the particle to be a preposition as part of a prepositional phrase with the following noun and apply the case that preposition governs (so [leysa af] + accusative becomes leysa + [af + dative]). Such is the case also in Figure 1 as the preposition fyrrir can govern either accusative or dative, but in this case the collocation gera raddi fyrrir (e. allow for, anticipate) exclusively takes dative.

Other typical errors involve the use of grammatical voice, and inflectional errors in closed word classes. Inflectional errors in nouns or verbs are among the most common errors as well but they are also prominent in the L1 corpus. Fixed word order in Icelandic is not intuitive for the learner either which created two additional error subclasses within syntax. Another very specific error type for L2 in the lexical category is context – an incorrect word chosen for the specific context often prompted by a literal dictionary translation.

The frequency of error codes was ranked to identify to which extent subclasses differ in frequency between the corpora. If the error code does not appear in a corpus, the rank is by default higher by one than the total number of ranks. The relative rank (Δ rank) between the corpora was calculated for each error code. A high number indicates a large difference in ranks between corpora for an error code, and a low number indicates similar rankings.

<table>
<thead>
<tr>
<th>Error Codes</th>
<th>Category</th>
<th>Rank L1</th>
<th>Rank L2</th>
<th>Δ rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>wording</td>
<td>wording</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>nonword</td>
<td>nonword</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>date-period</td>
<td>punctuation</td>
<td>99</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td>extra-conjunction</td>
<td>punctuation</td>
<td>32</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>comma4colon</td>
<td>punctuation</td>
<td>89</td>
<td>90</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Error codes with most similar rankings between the corpora

Table 3 shows the error codes that have the most similar frequency rankings in the L1 and L2 corpora. Two types of errors that are ranked among the highest in both corpora are wording and nonword error, the latter being possibly a simple typing error or an attempt to write a word form that does not exist, whereas the former is the most general error type which includes any type of formulating a phrase or a clause in a wrong way, and is often combined with other error types.

5 Conclusion

This paper introduces the Icelandic L2 Error Corpus, the first learner error corpus for Icelandic, which is a collection of texts written by users of Icelandic as a second language. The majority of the texts are essays submitted by students in the Icelandic as a second language program at the University of Iceland. The texts have been manually annotated for errors based on an error tagset previously built for the general Icelandic Error Corpus. Both corpora are openly accessible via CLARIN repository. First CIA results are also presented, comparing the L2 corpus with the general corpus.

At this point, the corpus consists of 12081 revision spans and 17508 categorized error instances. The preliminary results show a large disparity in the quantitative distribution of errors in the Icelandic L2 Error Corpus and the general Icelandic Error Corpus. This disparity relates to both the occurrence of different error categories, where grammar related errors are 4 times more prominent in the L2 corpus,
and the total error rate, which is 3 times as high for the L2 corpus compared to the native speaker referent. Moreover, it is still more than twice as high when the L2 speakers have reached the highest proficiency level.

Further analysis of the data is underway and we are in the process of compiling preliminary results of the L2 vs. L2 CIA. As the corpus is still small and the distribution of features such as first language is not as wide, the focus will be on the proficiency level and length of residence, which tend to intertwine. We hope to expand this corpus to provide further possibilities to analyse various features of learner language that are difficult to highlight due to the limited size of the sample. With the expansion of the corpus, it has potential to become an important asset for learning Icelandic.

References


Anton Karl ingason, Lilja Björk Stefánsdóttir, and Órunn Arnardóttir. 2020. Icelandic error corpus (IceEC) version 0.9. CLARIN-IS.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Órunn Arnardóttir, Xindan Xu, and Isidora Glišić. 2021. The Icelandic L2 error corpus (IceL2EC) version 1.1. CLARIN-IS.


