The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic

Isidora Glišić
University of Iceland
Reykjavik, Iceland
isg14@hi.is

Anton Karl Ingason
University of Iceland
Reykjavik, 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 consisting of 22,705 manually-annotated errors in different categories pertaining to grammar, spelling, lexical and other issues. The corpus was used to perform a contrastive interlanguage analysis, first using a native speaker reference corpus – the Icelandic Error Corpus, then analysing the corpus internally based on linguistic features relevant to second language acquisition. This paper presents the corpus and first results of the analysis.

1 Introduction

Icelandic is a small but increasingly popular language among language learners, both immigrants in Iceland trying to fit into the society and language enthusiasts at large. However, the popularity of Icelandic is a quite novel phenomenon and teaching materials are still scarce and constantly in development. With the rise of language technology efforts in Iceland, it is finally possible to utilize the new technologies in creating ICALL solutions and a major step towards this is creating an error corpus consisting of texts written by users of Icelandic as a second language. At the moment of writing, the Icelandic L2 Error Corpus is a collection of 85 texts, predominantly student essays, annotated for various types of errors. The corpus contains a total of 147,465 words, 15,571 revision spans and 22,705 error instances, where a revision span is a word or a continuous 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 likely to grow and is expected to 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. The theoretical background on learner corpora is discussed in section 2, followed by an overview of previous research on learner interlanguage for Icelandic and the introduction to the new Icelandic L2 corpus in section 2.1. Section 3 describes the methods that we used. Section 4 presents an analysis using two comparative methods.

2 Properties of L2 Mistakes and Error Corpora

The potential that learners' errors have as an indicator of the developmental stages they are likely to have reached in second language acquisition (in further text: SLA) has become obvious already in the 1960s (Thewissen, J., 2013). With the advancement of technology and the rise of corpus linguistics and computer-aided SLA since the early 1990s, more emphasis has been put on the importance of creating learner corpora. These are collections of texts that have been annotated for errors, as this provides access not only to the distribution of learner errors from various perspectives but to their entire interlanguage (Díaz-Negrillo, A., and Fernández-Domínguez, J., 2006), and is key in creating automatic correction tools, development of syllabi, curricula, exams, textbooks and graded readers for SLA.

Isidora Glišić and Anton Karl Ingason 2022. The Nature of Icelandic as a Second Language:An Insight from the Learner Error Corpus for Icelandic. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 23–33. DOI: https://doi.org/10.3384/9789179294441

Notable learner corpora include CITE https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

The first step in examining learner corpora is the standardization of error typologies. Most error tagging systems to date tend to be token-based and focus on five distinct linguistic error categories: non-word errors, grammatical errors, lexical errors, errors related to style, and punctuation errors. Non-word errors refer to simple spelling mistakes and accidental repeating of a character or word that result in a word that does not exist, and they are among the most common errors made by native speakers, as well as a specific type of context-related lexical errors which are commonly referred to as confusion sets (Golding, A., and Roth, D., 1999; Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S., 2009; Friðriksdóttir, S. R., and Ingason, A. K., 2020). These are often related to semantically distinct words that are homophones (e.g. *leyti* 'degree' and *leiti* 'hill' in Icelandic and *piece* and *peace* in English). However, for second language (henceforth, L2) users, grammatical and lexical errors are typically more prominent than in native speakers, and tend to decrease with advancing proficiency level, as the interlanguage is developing closer to the target language. For optimal analysis it is important to observe the size of the corpus and the diversity of submitted texts (both variety of authors and genres). As for the very process of language learning and errors that occur within the interlanguage, there are several factors that need to be considered, some of which are connected to the language situation or task (such as the genre and length of the text and use of reference tools) and others pertaining to the learner (their age, gender, proficiency level, mother tongue and other linguistic background) (Granger, S., 2008).

Scholars have been divided to what extent crosslinguistic influence plays a role. First language (L1) interference was considered crucial in SLA until more extensive research was conducted. Modern theories such as the processability theory emerged, which state that all language learners go through five distinct stages of grammar acquisition, regardless of their native language, and it is not possible to skip a stage or process them in a different order (Pienemann, M., 2011). The theory does not reject crosslinguistic interference but claims that only those linguistic forms that the learner can process can be transferred to the L2 (Pienemann, M., Di Biase, B., Kawaguchi, S., and Håkansson, G., 2005). Therefore, other second languages that the learner acquired before the target language are also relevant, as transference can occur from any other languages that the learner acquired and having internalized more than one grammatical system leads to a generally better understanding of language processing or meta-linguistic awareness (Cummins, J. 1991,). Nevertheless, relevant literature indicates that some classes of common errors are independent of native language background (Gamon, M., Leacock, C., Brockett, C., Dolan, W., Gao, J., Belenko, D., and Klementiev, A., 2013).

An important standard for assessing the stage of learners' interlanguage is the Common European Framework of Reference for Languages (CEFR) that was launched by the European Council in 2001 as an international standard for describing language ability. It describes language ability on a six-point proficiency scale - A1, A2, B1, B2, C1, C2. A is considered the beginner level, B1 intermediate, B2 advanced and C proficient (near-native) level (Piccardo, E., Goodier, T., and North, B., 2018).[1] The scale is particularly important in evaluating learner errors, as specific types of errors typically emerge on specific proficiency levels. Certain stagnation and regression points have been noted, particularly the one between B1 and B2, where the probability of some types of errors tends to increase rather than decrease. In this case, regression is viewed as a normal part of learning progress, as learners move towards more complex use of language and attempt making longer sentences (Thewissen, J., 2013).

The next section will review previous research on the acquisition of Icelandic as L2 and introduce the L2 error corpus for Icelandic, a novel kind of resource in the context of Icelandic language technology that is still in development.

2.1 Resources for Studying L2 Errors in Icelandic

To understand the context of teaching Icelandic as a second language, one must bear in mind that Icelandic is a small language that was historically spoken by a homogeneous population. For a very long

[1] For more details about proficiency level assessment scale, see: https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

time, not many foreigners were interested in learning this language and no textbooks or teaching methodology existed. It was not until the 1980s when Svavar Sigmundsson, using contrastive linguistics methods, decided to analyze mistakes of learners of Icelandic, predominantly those of Scandinavian origin (Sigmundsson, S., 1987). As the interest in learning Icelandic as a second language grew, the first textbooks started being published and teaching methodology started developing, setting the standard for the order of grammar acquisition for Icelandic later reiterated and revised many times. However, it was not until very recently that attention was drawn to learner errors in adopting the syllabus to the natural order of acquisition. Using the processability theory, Sigríður Þorvaldsdóttir and María Garðarsdóttir (Þorvaldsdóttir, S., and Garðarsdóttir, M., 2013) started looking into the order of acquisition of cases for their learners‘ interlanguage on the lowest proficiency levels. Most recently Gísli Hvanndal Ólafsson has examined the general acquisition of grammar from absolute beginners to level A1 (Ólafsson, G. H., 2016) – including cases, verb conjugations and declension of nouns, adjectives and pronouns. Finally the learner corpus was published last year (Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I., 2021) and is at this point still in development. The corpus in its current form is published in a CLARIN repository under a CC BY 4 license.

The Icelandic L2 Error Corpus[2] currently consists of 85 texts from 36 second language speakers of Icelandic with 15 different first languages, containing 22,705 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 under their name or anonymously. The call for texts was first directed to the students of Icelandic as a second language at the University of Iceland but was subsequently extended to a public call. At the moment of writing, 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 changes drastically when separated by skill level (with the mean for A1 texts being 324 words and 5177 words for C2) as both the written language expression ability and the nature and type of the texts vary - the highest skill level texts typically being long academic essays and parts of or entire MA theses. For more numerical data based on skill level, see Table 4. The currently small size of the corpus (however relatively large when compared with other learner corpora and taking into account the number and variety or the annotated errors) and slow process of acquiring texts are related to the protection of authors' rights, as the University and language school do not have authority to share students' essays and the authors need to submit texts and fill out the publication agreement themselves.

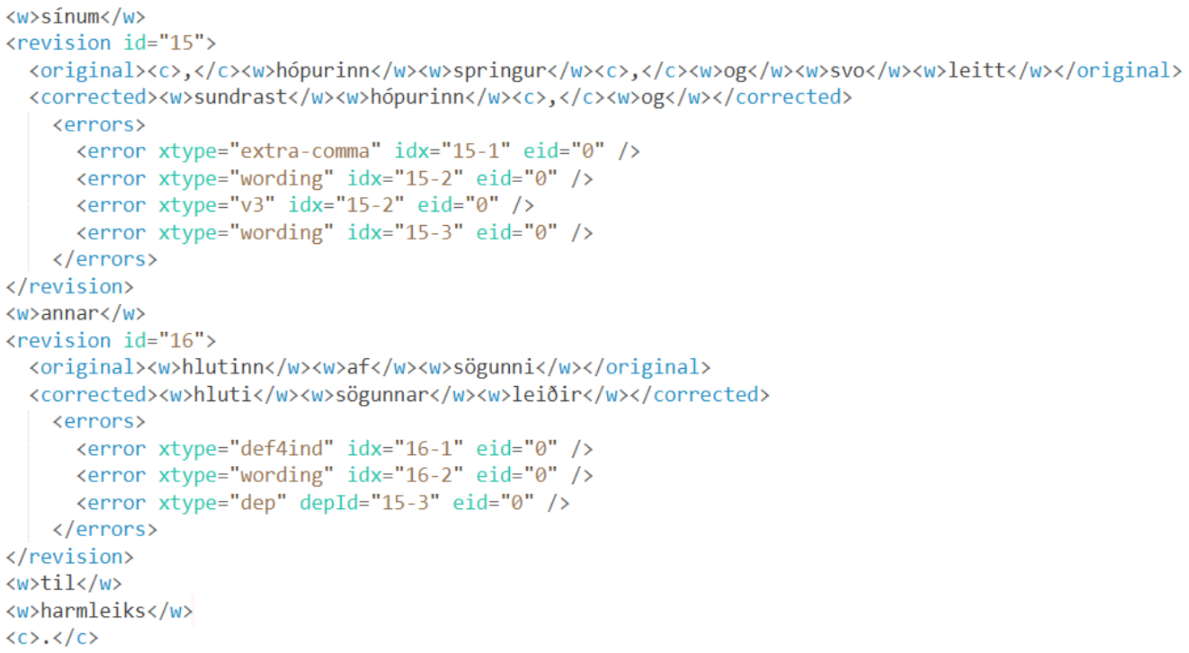
The advantage of using student essays is the accessibility of texts (as it is otherwise very difficult to obtain texts in Icelandic written by foreigners) from subjects with different first and second language background. Furthermore, it is also relatively easy to estimate their proficiency level based on their study progress. 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, M., and Þorvaldsdóttir, S., 2020). However, due to the nature of the writings (academic texts) some types of errors tend to be more prominent than in other types of writings. Apart from that, many generic errors might be removed as the texts tend to be polished for better academic success. Texts that arrived from outside of the University were separately scrutinized and the proficiency level was estimated based on the CEFR scale. Bearing in mind the relevant factors in SLA mentioned in section 2, other required information that the subjects provided themselves include their native language, second language(s), length of residence in Iceland and how long they have been learning Icelandic at the time of writing of the submitted text. Other basic demographic information such as age and gender are also part of the form, but are not required. As a relevant number of subjects chose to omit this data, it was not taken into the analysis.

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

[2] The corpus is available at: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/106

3 Methods

The texts for the Icelandic L2 Error Corpus were collected through an open online publication agreement and manually proofread and mapped for errors. Microsoft Word's track changes feature was used for this because it preserves the original version of the text along with the corrected version. After the proofreading process, 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 complex revision span containing several error codes and a dependent error.



```
<w>sínum</w>
<revision id="15">
  <original><c>,</c><w>hópurinn</w><w>springur</w><c>,</c><w>og</w><w>svo</w><w>leitt</w></original>
  <corrected><w>sundrast</w><w>hópurinn</w><c>,</c><w>og</w></corrected>
    <errors>
      <error xtype="extra-comma" idx="15-1" eid="0" />
      <error xtype="wording" idx="15-2" eid="0" />
      <error xtype="v3" idx="15-2" eid="0" />
      <error xtype="wording" idx="15-3" eid="0" />
    </errors>
</revision>
<w>annar</w>
<revision id="16">
  <original><w>hlutinn</w><w>af</w><w>sögunni</w></original>
  <corrected><w>hluti</w><w>sögunnar</w><w>leiðir</w></corrected>
    <errors>
      <error xtype="def4ind" idx="16-1" eid="0" />
      <error xtype="wording" idx="16-2" eid="0" />
      <error xtype="dep" depId="15-3" eid="0" />
    </errors>
</revision>
<w>til</w>
<w>harmleiks</w>
<c>.</c>
```

Figure 1: An example of revision spans with multiple error codes and a dependent error.

The figure 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 error id "15-1" refers to the first character in the revision span, whereas the two words that follow need to be covered by two different error types as there is both invalid syntax (permutation of word order) and a erroneous choice of words involving both of them and both error codes are labeled "15-2". The error code *dep* is not included in the annotation list as its purpose is to annotate that the error in question is a dependent error connected to another in a different revision span, using the original error's *idx*. The corpus syntax along with the error annotation system used for error labeling was originally developed and used for the Icelandic Error Corpus (Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, Þ., 2020) which contains errors in native speaker texts. However, as the applicability of the corpora extended and they are being used in creating a spelling and grammar correction package for Icelandic,[3] it became evident that relying on this formatting of revisions spans is sub-optimal, as it makes it impossible to know automatically which subset of the tokens of the revision spans are concerned with which errors. The spans are therefore being revised in the new version of the corpora and should ultimately include information that should connect each word in the span to a specific error it is related with.

The annotation system that was originally created has also undergone changes in the process of creating the L2 corpus, as new labels needed to be added for errors that were specific for second language use (the list of errors specific to the L2 corpus can be viewed in Table 2). The error tagset consists of 6

[3] The correction package, *Greynir Correct*, is available at: https://github.com/mideind/GreynirCorrect

main categories (*coherence*, *grammar*, *orthography*, *style*, *vocabulary*, *other*) which are further divided into subcategories. Some subcategories are very narrow while others are more wide-ranging (notably the Orthography-Punctuation category) and in total there are 259 error codes. The codes are meant to be short but descriptive and are often abbreviated versions of the issue they pertain to, e.g. *simple4cont* stands for simple-for-continuos, where a verb is used in the simple tense and should be in the continuous tense; *agreement-pred* signifies that a predicate is not in agreement with its subject. Some error codes are more specific, such as *geta*, which indicates that the Icelandic auxiliary verb 'geta' (e.'be able to') is wrongly used with an infinitive or present tense instead of past participle. A list of all the codes along with an example and a description is available at `https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/errorCodes.tsv`.

After the dataset of TEI documents has been finalized (note that new texts are still likely to be added and the corpus is a work in progress), statistical analyses were conducted that included quantifying the number of texts, revision spans and error occurrences in the corpus, 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 and will be analysed in the next section.

4 Data Analysis

As stated before, learner corpora can provide invaluable insight into the learners' interlanguage, uncovering various linguistic features depending on the variables that the analysis focuses on. The method primarily used for this purpose 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, S., 2008). The former can uncover the distinguishing features of L2 language use while the second 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 finalized, this provides us with the possibility to make a CIA based on the first type mentioned, and the results will be presented in the following section. For the L2 vs. L2 analysis, as the corpus is still small and the distribution of features such as age or mother tongue is not as wide, the focus will be on the proficiency level and length of residence, which tend to intertwine.

4.1 General Characteristics of L2 Errors Comparative to L1 Errors

Corpus	Files	Total words	Revisions	Categorized Errors	Errors/1000w
Icelandic Error Corpus	4,046	1,137,941	44,261	55,346	44.56
Icelandic L2 Error Corpus	85	147,465	15,571	22,705	153.97

Table 1: Numerical data for both L1 and L2 Icelandic error corpora.

To compare the errors of L2 speakers to native speakers, a contrastive analysis was conducted between the L2 error corpus and the general Icelandic Error Corpus (Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, Þ., 2020).

As Table 1 demonstrates, the number of errors per 1000 words is significantly higher in L2 texts than in the general corpus, and despite the general corpus being much larger with tenfold total word count, the total number of errors in the L2 corpus is still quite statistically significant.

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 so this may not be reflective of L2 users as a population.

The analysis also sheds light on a significant disparity in frequency of certain error categories and subcategories in L2 Icelandic compared to L1 errors. The most frequent error category in the L2 corpus is grammar, which accounts for almost half of all errors (43.57%). In comparison, the category grammar accounts for only 11.8% in the general Icelandic Error Corpus. The punctuation (12.14%) and wording (11.63%) subcategories are most prominent in second tier, where *wording* is also a specific error code with highest frequency (unsurprisingly, as many otherwise unsorted errors connected to choice of words tend to fall under it). Each other error category comprises 5% or less of total errors. Depicted in Table 2 are all error codes that appear only in the L2 corpus, 30 of which are within the grammar category.

Grammar	
case-verb	case-prep
genitive	pro-inflection
act4mid	missing-sub
tense4perfect	missing-fin-verb
case-collocation	act4pass
extra-sub	mid4act
extra-dem-pro	missing-dem-pro
v3-subordinate	numeral-inflection
perfect4tense	missing-obj
adj4noun	noun4adj
mid4pass	case-adj
pass4mid	pass4act
passive	geta
extra-fin-verb	extra-prep
extra-munu	syntax-other
Orthography	
wrong-symbol	
Vocabulary	
context	interr-pro
though	þar4það

Table 2: Error codes that appear only in the L2 error corpus

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, as well as the use of grammatical voice, and inflectional errors in closed word classes. Inflectional errors in nouns or verbs are also among the most common errors but 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. When the frequencies of multiple error codes were identical, they were ranked equally. 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.

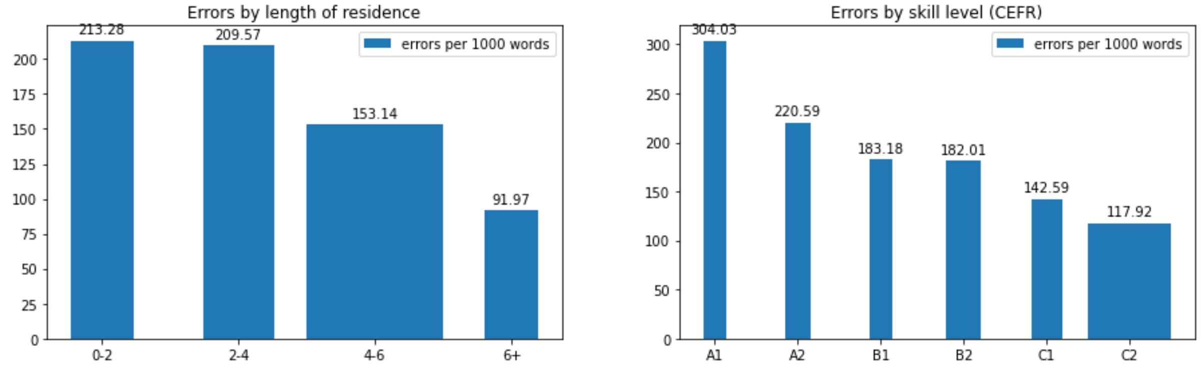
Table 3 shows that the two types of error that are ranked among highest in both corpora are *wording* and *nonword* error, the latter being possibly a simple typing error or an attempt to write a word form that

Error Codes	Main Category	Subcategory	Rank L1	Rank L2	Δ rank
wording	style	wording	1	1	0
nonword	orthography	nonword	3	3	0
date-abbreviation	orthography	punctuation	99	99	0
extra-conjunction	vocabulary	insertion	32	33	1
comma4colon	orthography	punctuation	89	90	1

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

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.

4.2 Errors by Proficiency Level and Length of Residence



(a)

(b)

Figure 2: Errors per 1000 words based on length of residence and proficiency level.

Factors that are generally considered in evaluating the interlanguage development are the length of study of the target language as well as the level of interaction and use of the language (does the learner live in the country where the language is spoken, how much are they exposed to the language daily and through which outlets). However, recent research has advised against relying on criteria that assign proficiency by length of study in a language learning program and suggest rather that each learner's production be individually assessed (Thewissen, J., 2013).

Level	**Files**	**Total words**	**Total errors**	**Errors/1000w**
A1	19	7,759	2,359	304.03
A2	19	11,900	2,625	220.59
B1	12	12,900	2,363	183.18
B2	11	19,504	3,550	182.01
C1	10	22,617	3,225	142.59
C2	14	72,785	8,583	117.92

Table 4: Total number of files, words, errors and errors per 1000 words per skill level.

In this corpus, the proficiency level that learners achieve mostly correlates with the time they have spent residing in Iceland. For 31 of the submitted texts, the author started learning the language before arriving to the country, whereas for the remaining texts the learner started learning the language immediately or some time after starting to live in Iceland and is likely to have not had much contact with the language before commencing the formal learning process. The two graphs in Figure 2 show the number of errors per 1000 words based on length of residence and the proficiency level and show a downward curve which was to be expected as SLA progresses. The width of the bars indicates the total number of words within the given category, notably highest in 4-6 years of residence and level C2, as was briefly mentioned in section 2.1 and laid out in Table 4 (as received texts on this level were mostly longer essays or entire theses). An interesting development is that the downward trend is not as sharp between levels B1 and B2, and in fact the frequency of *wording* and *nominal-inflection* errors increases, which is in line with the typical regression point mentioned in section 2, that the probability of some types of errors increases rather than decrease between levels B1 and B2. Even though this type of regression is expected and observed through SLA research, it needs to be stated that there are no strict guidelines and defined grammatical and lexical requirements that learners need to reach to officially be on a certain level on the CEFR scale for Icelandic. Therefore, it is entirely possible that at the beginning of their second year of the BA in Icelandic as a second language, not all students have reached level B2 and there is likely some overlap in the labelling of texts between these two CEFR levels.

Finally, we should also keep in mind that the general error corpus has the average of 44 errors per 1000 words whereas the average for the highest level (C2) and longest dwelling (more than 6 years) is 117 and 91 respectively which shows that non-native speakers are more than twice as likely to make mistakes even as they approach near-native competence as much as possible. However, the nature of errors changes over time.

Table 5 shows the frequency of the (5) most common errors per proficiency level. *Nominal-inflection* is an error type that consistently ranks among the most frequent. This is not surprising, and as Þorvaldsdóttir and Garðarsdóttir point out in their research (Garðarsdóttir, M., and Þorvaldsdóttir, S., 2020), acquisition of cases is a slow process with many typical points of overgeneralization (subject as nominative, object as accusative, atypical subject as dative etc.), and the results indicate that the learners assume the so-called structural case later than thematic case with the most atypical idiosyncratic case being the latest and most reluctantly accepted which in most cases is the genitive case. A common source of case errors also comes from the previously mentioned case governance which also ranks high on all levels. For example, L2 users will commonly misuse a phrasal verb and instead of interpreting it as a verb clause they would take the preposition as part of a prepositional clause with the following noun and apply the case that preposition governs (so [*leysa af*] + accusative becomes *leysa* +[*af* + dative]).

Although the general wrong-choice-of-words error, *wording*, ranks consistently highest (and as was shown, is also the most frequent error type in the general corpus), the occurrence of the more particular choice of word error, *context*, drops between levels, being in 4th place on A2 and dropping to the 7th place on level C1. Another common error type, *nonword*, is likely to not be a competence error but accidental or caused by over- or underuse of specific Icelandic accented vowels. However, its frequency being significantly higher on the lowest proficiency level does to some extent stem from overgeneralization, e.g. assigning a wrong gender to a noun creating an incorrect inflection form, or conjugating an irregular verb as regular. Other noteworthy error types are *ind4sub* / *sub4ind* which track the incorrect use of the subjunctive mood and rise in frequency as proficiency level increases (particularly *ind4sub*, as learners tend to overuse the indicative mood). This is not surprising, as beginner and lower intermediate learners use simpler sentence structures and do not learn about grammatical mood until later on in the learning process, and this type of grammatical error is among the most common for native speakers as well.

Lastly, there is a number of ambiguous cases where it is not clear whether an error is a spelling or a grammatical error and in these cases it is hard to estimate the author's intention. Here the error would be categorized based on the overall analysis of the given text and the skill level and its linguistic expectations, i.e. if the text has repeated unambiguously inflectional errors, the error in question would most likely be categorized as such, whereas if the grammatical correctness of the text overall is high, it would be interpreted as a spelling-orthography related error.

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 student 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 based on native speaker texts. First results of two CIA approaches are also presented, first comparing the L2 corpus with the general corpus and second analysing the L2 error corpus focusing on proficiency level.

Error Codes	Category	Subcategory	Freq	Errors/1000w
A1				
wording	style	wording	236	30.42
nominal-inflection	grammar	inflection	115	14.82
nonword	orthography	nonword	91	11.73
missing-period	orthography	punctuation	76	9.79
extra-word	vocabulary	insertion	72	9.28
A2				
wording	style	wording	260	21.85
nominal-inflection	grammar	inflection	185	15.55
wrong-prep	grammar	prep	97	8.15
context	vocabulary	lexical	91	7.65
extra-comma	orthography	punctuation	88	7.39
B1				
wording	style	wording	254	19.69
nominal-inflection	grammar	inflection	93	7.2
extra-word	vocabulary	insertion	83	6.43
extra-comma	orthography	punctuation	83	6.43
def4ind	grammar	definitiveness	81	6.27
B2				
wording	style	wording	558	28.61
nominal-inflection	grammar	inflection	202	10.36
extra-word	vocabulary	insertion	87	4.46
nonword	orthography	nonword	84	4.31
missing-word	vocabulary	omission	84	4.31
C1				
wording	style	wording	453	20
nonword	orthography	nonword	130	5.75
ind4def	grammar	definitiveness	108	4.77
nominal-inflection	grammar	inflection	105	4.64
agreement-concord	grammar	agreement	99	4.38
C2				
wording	style	wording	802	11.02
nominal-inflection	grammar	inflection	508	6.98
nonword	orthography	nonword	309	4.24
wrong-prep	grammar	prep	296	4.07
extra-comma	orthography	punctuation	249	3.42

Table 5: Most common error codes by proficiency level in the L2 corpus

At this point, the corpus consists of 15,571 revision spans and 22,705 categorized error instances. When compared to the L1 error corpus for Icelandic which has more than a million words and 44,261 revision spans, the overall number of errors is only slightly less than half of the number of errors in the general corpus, which is both valuable for research and analysis and developing the correction tool, and indicates a great distinction in the frequency of errors L2 and L1 users make in written form, as was further shown in our analysis. It should be noted that a total of 85 texts by 36 learners with only 15 first languages, and at the moment of writing, the immigrant population in Iceland uses more than 100 different native languages, might not be fully representative of the L2 community in Iceland (and we hope the further expansion of the corpus will provide more diversity).

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 and dwelled in the country for more than 6 years. The L2 vs. L2 analysis also yielded interesting yet predictable results, showing the downwards trend of error occurrences as the learner's proficiency and their length of residence in Iceland increases, with a notable slight increase in certain types of errors between levels B1 and B2.

Learner error corpora are important for shedding light on learner interlanguage which can aid the development of various automatic language correction tools and teaching materials and also provide insight into how and in which order certain grammatical and lexical categories are acquired and internalized. Thus we hope to further expand this corpus to provide more possibilities to analyse various features of learner language that could not be covered so far due to the limited size of the sample and its lack of diversity of highlighted linguistic features. With the expansion of the corpus, it has potential to become an important asset for learning Icelandic.

References

Cummins, J. 1991. *Interdependence of first- and second-language proficiency in bilingual children*, page 70–89. Cambridge University Press.

Díaz-Negrillo, A., and Fernández-Domínguez, J. 2006. Error tagging systems for learner corpora. *Revista española de lingüística aplicada, ISSN 0213-2028, Vol. 19, 2006, pags. 83-102*, 19, 01.

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

Gamon, M., Leacock, C., Brockett, C., Dolan, W., Gao, J., Belenko, D., and Klementiev, A. 2013. Using statistical techniques and web search to correct ESL errors. *CALICO Journal*, 26:491–511, 01.

Garðarsdóttir, M., and Þorvaldsdóttir, S. 2020. A processability approach to the development of case in L2 Icelandic. *Language, Interaction and Acquisition A cross-theoretical and cross-linguistic perspective on the L2 acquisition of case systems / Lacquisition de systèmes casuels en L2 : des études à travers plusieurs théories et langues*, 11(1):68–98.

Golding, A., and Roth, D. 1999. A winnow-based approach to context-sensitive spelling correction. *Machine learning*, 34(1-3):107–130.

Granger, S., 2008. *Learner Corpora in Foreign Language Education*, pages 1427–1441. 01.

Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. 2009. Context-sensitive spelling correction and rich morphology. In Jokinen, K. and Bick, E., editor, *Proceedings of NODALIDA 2009*, pages 231–234.

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I. 2021. The Icelandic L2 error corpus (IceL2EC) version 1.1. CLARIN-IS.

Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, Þ. 2020. The Icelandic error corpus (IceEC). version 1.0.

Piccardo, E., Goodier, T., and North, B. 2018. *Council of Europe (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Strasbourg: Council of Europe Publishing., 01.

Pienemann, M., Di Biase, B., Kawaguchi, S., and Håkansson, G. 2005. Processing constraints on L1 transfer. *Handbook of bilingualism: Psycholinguistic approaches*, pages 128–153, 01.

Pienemann, M. 2011. *Studying processability theory : an introductory textbook / edited by Manfred Pienemann, Jorg-U. Kessler.* Processability approaches to language acquisition research teaching (PALART), v. 1. John Benjamins Pub. Co., Amsterdam ;.

Sigmundsson, S. 1987. Íslenska í samanburði við önnur mál. *Íslenskt mál og almenn málfræði*, 9.

Thewissen, J. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1):77–101.

Ólafsson, G. H. 2016. Grammar and linguistic structures at level A1 of Icelandic. Master's thesis, University of Iceland, Unpublished, 6.

Þorvaldsdóttir, S., and Garðarsdóttir, M. 2013. Fallatileinkun í íslensku sem öðru máli. *Milli mála: Tímarit um erlend tungumál og menningu*, 5:45–73.