Error Corpora for Different Informant Groups: Annotating and Analyzing Texts from L2 Speakers, People with Dyslexia and Children

Þórunn Arnardóttir, Isidora Glišić, Annika Simonsen,
Lilja Björk Stefánsdóttir, Anton Karl Ingason
University of Iceland
Sæmundargata 2, 102 Reykjavík, Iceland
{thar, isg14, ans72, lbs, antoni}@hi.is

Abstract

Error corpora are useful for many tasks, in particular for developing spell and grammar checking software and teaching material and tools. We present and compare three specialized Icelandic error corpora; the Icelandic L2 Error Corpus, the Icelandic Dyslexia Error Corpus, and the Icelandic Child Language Error Corpus. Each corpus contains texts written by speakers of a particular group; L2 speakers of Icelandic, people with dyslexia, and children aged 10 to 15. The corpora shed light on errors made by these groups and their frequencies, and all errors are manually labeled according to an annotation scheme. The corpora vary in size, consisting of errors ranging from 7,817 to 24,948, and are published under a CC BY 4.0 license. In this paper, we describe the corpora and their annotation scheme, and draw comparisons between their errors and their frequencies.

1 Introduction

Error analysis is a crucial part of corpus linguistics and applied linguistics as it provides an insight into language use and the needs of speaker groups within a language. It thereby facilitates the development of a variety of practical tools to aid these needs, such as more focused teaching and learning materials, and software tools like spelling and grammar checkers. To contribute to this field, we present three Icelandic error corpora, each focusing on different speaker populations; second language users of Icelandic (hereinafter: L2 Icelandic), children at the age of 10 to 15, and people with dyslexia. Thus, we have created three manually annotated error corpora of different sizes, one for each respective informant group, and extracted statistical data on the errors that occur. These corpora are an invaluable source for further research in Icelandic for both academic and practical purposes. All corpora are published under a CC BY 4.0 license. (Ingason et al., 2022b,a, 2021)

The paper is structured as follows. Section 2 discusses the creation of error corpora in general. Section 3 describes the specialized error corpora, their annotation and the three respective corpora. Section 4 draws comparisons between the errors in the three specialized corpora and compares them to errors in a previous general error corpus. Section 5 discusses possible future use of the corpora and finally, we conclude with Section 6.

2 Creating Error Corpora

Error analysis has been an integrated part of applied linguistics and computational linguistics for decades, and corpus linguistics in general has developed as a key methodology in the humanities and social sciences (Paquot and Gries, 2020). It provides key insight into both the errors that adult native speakers of a language produce in writing, as well as those of language learners, children, and people with different learning difficulties such as dyslexia. Gathering data on these errors has become a standard practice for many languages and is invaluable for creating different software tools for language correction and suggestion, such as spell checkers, grammar assistance, and lexical and stylistic suggestions. Furthermore, an error corpus gives way to contrastive analysis which leads to better understanding of language use in different groups and the creation of both digital and analogue content that would facilitate them (e.g. improving teaching materials for second language learners and children).

The Icelandic Error Corpus was created for this purpose, and was the first Icelandic error corpus (Arnardóttir et al., 2021). It has already been used for developing an Icelandic open-source spell and grammar checker (Óladóttir et al., 2022), wherein the labeled errors in the corpus are used to measure the spell and grammar checker's improvements. This error corpus consists of texts written by Icelandic native-speaking adults with no known learn-

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 245 - 252

ing disabilities, and provides information on errors which this speaker group is likely to make. However, the Icelandic population consists of various speakers who might make errors different to a general speaker. For this purpose, error corpora for particular speaker groups are important. Analyzing errors made by these groups enables the development of spell and grammar checkers and practical tools suited for those groups, as well as facilitating effective teaching methods and materials.

3 The Icelandic Specialized Error Corpora

Three Icelandic error corpora were created between the fall of 2019 and the fall of 2022, reflecting three different user groups; The Icelandic L2 Error Corpus, The Icelandic Dyslexia Error Corpus, and The Icelandic Child Language Error Corpus. All corpora are published under a CC BY 4.0 license in the Icelandic CLARIN repository (Ingason et al., 2022b,a, 2021). An older version of the Icelandic L2 Error Corpus was described in Glisic and Ingason (2022).

3.1 Annotation

The Icelandic Specialized Error Corpora all have the same annotation scheme and structure, which is shared by the Icelandic Error Corpus (Arnardóttir et al., 2021). The steps involved in creating the corpora were gathering large quantities of texts within each focus category, manually proofreading the texts for errors, and finally creating the corpus in the decided digital format. Each error in the texts was then manually labeled within a pre-decided annotation scheme. The corpora are published in augmented TEI-format XML documents, making them machine readable so that corpus management platforms particular to TEI format files can be used to obtain information from the corpus. A specific TEI element, *revision*, was created to map out the differences between the original text file and the manually corrected file. Each XML document consists of many revision spans that include the mismatching text and one or more error tags that are manually classified within a previously decided annotation scheme.

Errors in the original texts were detected by following Icelandic spelling and grammar rules. Many of these rules are included in the Icelandic language council's spelling rules.[1] Rules on language usage are included in a resource called *Málfarsbankinn*[2] (direct translation: *The Language Usage Bank*). This is a collection of rules and general advice concerning grammar, fixed phrases, spelling, and more. In addition to these explicit errors, stylistic errors were also corrected, i.e. errors which are not included in the aforementioned resources, but belong to known guidelines for writing text. These errors for example include using numerals instead of numbers in particular cases.

Language error classification can be done in many different ways, but two major categories are mostly defined as linguistic errors (morphology, syntax, etc.) and surface structure taxonomies (omission, addition, etc.), where most studies combine the analysis of both these categories (Macdonald et al., 2013). This practice was adopted for the Icelandic error corpora and a particular annotation scheme was created. It evolved as the error annotation progressed and new types of texts came in — particularly many new error categories were noted with L2 texts (more on this in Subsection 3.2).

The annotation scheme is hierarchical with three layers. Errors are classified within five main categories: orthography, grammar, vocabulary, coherence, and style. Each main category is further divided into more descriptive subcategories, which are then divided into error codes, 258 in total.[3] Each error code describes a specific type of error, although the scope and particularity can vary. For example, the code 'af4að' is used when the preposition *að* 'to' is mistakenly replaced by *af* 'of', whereas 'wrong-prep' is used in general with incorrect prepositions. 'i4y' is used when letters "i" and "y" are mixed up in a word but 'letter-rep' is used when a letter incorrectly replaces another one. This difference in scope is because both initial analysis and previous research on e.g. learner language indicated that certain specific errors occurred quite frequently, and the more detailed the annotation system, the better insight we can have into these errors, which will be of great value for future research. 'Wording' is the most general error type, and includes any type of formulating a phrase or a clause in a wrong way. Finally, some detected errors are connected to another error(s) within the sentence, such as in 'wording' and errors connected

[1] https://ritreglur.arnastofnun.is

[2] http://malfar.arnastofnun.is

[3] The annotation scheme is accessible at https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/errorCodes.tsv.

to syntax. These are classified within a separate category, 'other', which includes only one error code, 'dep', representing a dependent error.

Initial work on the Icelandic error corpus started in the autumn of 2019, and as of January 2020, several proofreaders were working with the texts; at one point a total of 12 people were reading over and correcting. Five specialists (either language technologists or Icelandic language specialists) worked on converting the texts into corpus data, creating the annotation system, and finally categorizing the errors found in the texts. Texts in the specialized error corpora needed to be collected from private sources, since no freely accessible texts written by these user groups were available. This proved to be a difficult and time-consuming process, because awareness about the project had to be raised within the interest groups and they had to be encouraged to participate. Interested authors signed publication agreements, which differed between user groups, and is discussed further in the sections pertaining to each corpus.

3.2 The Icelandic L2 Error Corpus

Icelandic is an increasingly popular language among language learners; there is a sizable population of immigrants in Iceland, who learn Icelandic to integrate into society. Additionally, there are people who are interested in learning Icelandic because they are language enthusiasts. However, teaching materials for Icelandic as a second language are scarce and in high demand. The creation of an L2 error corpus is a major step towards facilitating better teaching materials and also language learning tools (Glisic and Ingason, 2022). The version of the Icelandic L2 Error Corpus which is discussed here is an improved version of the one discussed in Glisic and Ingason (2022). More data has been added, and as a result, more errors have been collected.

The Icelandic L2 Error Corpus is a collection of 101 texts, predominantly student essays, written by 44 non-native speakers of Icelandic with 17 different native languages, containing in total 24,948 error instances in 17,241 revisions. Table 1 shows this, along with word count and frequency of errors per 1,000 words, which is 153.93.

Figure 1 and Table 2 depict the error rate per 1,000 words based on skill level, where the width of the bars indicates the number of words submitted for each level. Skill levels are shown according to the Common European Framework of Reference for Languages (CEFR), which is 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 (North and Piccardo, 2020). As mentioned, the corpus as a whole has 153.93 errors per 1,000 words, but the number of errors varies based on the authors' accomplished skill level and steadily drops in accordance with the language learning progress.

Revisions	Errors	Files	Words	Errors/1,000w
17,241	24,948	101	162,071	153.93

Table 1: Number of revisions, errors, files, words, and errors per 1,000 words in the Icelandic L2 Error Corpus

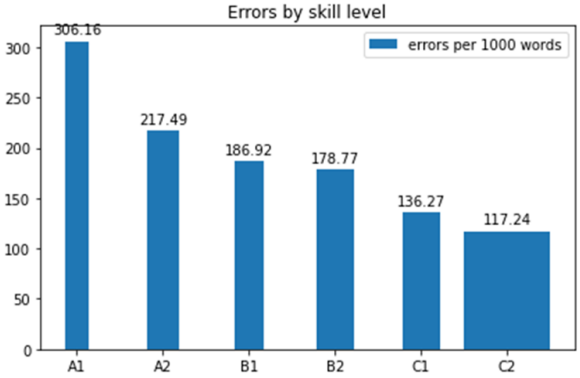

Figure 1: Error rate per 1,000 words in the Icelandic L2 Error Corpus according to learner level.

Level	Files	Total words	Total errors	Errors/1,000w
A1	20	7,960	2,437	306.16
A2	22	14,695	3,196	217.49
B1	14	16,071	3,196	186.92
B2	14	21,447	3,834	178.77
C1	16	27,871	3,798	136.27
C2	15	74,027	8,679	117.24

Table 2: Number of files, words, errors, and errors per 1,000 words in the Icelandic L2 Error corpus according to proficiency level.

In Table 3, we display the top 10 most common errors in the Icelandic L2 Error Corpus. The most common error is 'wording', which makes up 10.96% of the errors. 'Punctuation' and 'inflection' follow closely behind at 9.60% and 9.04%. Bear in mind that here we include L2 speakers of all proficiency levels. It is interesting to note that while 'inflection' is the third most common error in the L2 corpus, it is not even in the top 10 most common

errors in the Icelandic Error Corpus (Arnardóttir et al., 2021).

Subcategory	Main category	n	prop
wording	style	2735	11.0
punctuation	orthography	2396	9.6
inflection	grammar	2256	9.0
miscellaneous	other	1895	7.6
agreement	grammar	1526	6.1
prep	grammar	1452	5.8
definitiveness	grammar	1186	4.8
typo	orthography	1153	4.6
syntax	grammar	1146	4.6
insertion	vocabulary	1133	4.5

Table 3: Top 10 most frequent errors in the Icelandic L2 Error Corpus according to subcategory.

As mentioned in Subsection 3.1, the texts were previously unpublished and obtained directly from their authors. The text collection effort lasted from September 2020 to May 2022. 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. As a result of the call to students at the university, the texts are mainly student essays submitted for evaluation in various courses at the University of Iceland. Authors signed publication agreements, wherein they stated their native language, agreed to the text being published, and could choose to be anonymous or not. Out of 44 authors, seven chose to not be anonymous.

3.3 The Icelandic Dyslexia Error Corpus

Dyslexia is a learning disability, causing difficulty with reading but also with writing. In general, people with dyslexia make more misspellings than people who do not have it, and also different types of errors, which can affect how useful general spell and grammar checkers are for people with dyslexia. This difference in error rate and error type has not been previously studied for Icelandic informants, and for that reason, the Icelandic Dyslexia Error Corpus was created.

The dyslexia texts were collected through an open call. The only criteria for the texts was that their authors' native language was Icelandic, and that they had been diagnosed with dyslexia, but no proof of a diagnosis was asked for. The texts were written by informants born between 1961 and 2004 and some texts were written by the same author. For ethical reasons, as dyslexia is a medically diagnosed disorder, no more information on the authors was retained.

The Icelandic Dyslexia Error Corpus consists of 35 files totaling 38,891 words, and has 5,075 revisions with 8,436 errors (see Table 4). The error rate in the Icelandic Dyslexia Error Corpus is 216.91 errors per 1,000 words, which is the highest error rate of all the corpora, exceeded only by the A1- and A2-level L2 speakers.

Revisions	**Errors**	**Files**	**Words**	**Errors/1,000w**
5,075	8,436	35	38,891	216.19

Table 4: Number of revisions, errors, files, words, and errors per 1,000 words in the General Dyslexia Error Corpus.

Table 5 displays the 10 most frequent errors in the Icelandic Dyslexia Error Corpus, according to subcategory. The corpus shares similarities with the Icelandic Error Corpus in that among the most common errors are 'punctuation' and 'wording', but unlike the other corpora, the dyslexia corpus has a higher proportion of typos; a characteristic which is expected of people with dyslexia. The 'nonword' error code (a non-compound that does not appear in the dictionary) is also relatively high compared to the other corpora.

Subcategory	Main category	n	prop
typo	orthography	905	10.7
punctuation	orthography	903	10.7
wording	style	812	9.6
nonword	orthography	758	9.0
miscellaneous	other	545	6.5
syntax	grammar	524	6.2
spacing	orthography	443	5.2
insertion	vocabulary	411	4.9
omission	vocabulary	361	4.3
spelling	orthography	355	4.2

Table 5: Top 10 most frequent errors in the Icelandic Dyslexia Error Corpus according to subcategory.

Texts included in this corpus were collected over a two-year period, between October 2020 and October 2022, by different means. A collaboration with the Icelandic Dyslexia Association and school counselors at Icelandic colleges was established, and an open call to people with dyslexia was sent out. Authors who submitted their texts signed a

publication agreement, wherein they confirmed that they had been diagnosed with dyslexia and agreed that their texts would be published. Additionally, they could choose to be anonymous or not, but only two authors chose not to be anonymous.

3.4 The Icelandic Child Language Error Corpus

The Icelandic Child Language Error Corpus consists of 119 texts written by children aged 10 to 15 (born between years 2005 and 2010). The corpus excludes texts which were written by children with dyslexia as well as any text written by a child whose first language is not Icelandic. The interest in children's texts arose from the need to gain plausible insight into their vocabulary and grammar use, and the struggles they face in the process of language acquisition and learning to write. So far, we have been aware that children do not make the same mistakes as adults in their writing, based on tentative assumptions, teachers' experience and some studies of child language from other languages. Creating this corpus provides a unique opportunity to map out the exact errors and apply the findings directly in facilitating language and literacy development on elementary school level.

The corpus contains 7,817 errors, with an error rate of 208.77 per 1,000 words, as seen in Table 6. This is the second highest overall error rate in the four corpora, after the dyslexia corpus, although the L2 informants at proficiency level A1 and A2 have the highest error rates.

Revisions	Errors	Files	Words	Errors/1,000w
5,079	7,817	119	37,443	208.77

Table 6: Number of revisions, errors, files, words, and errors per 1,000 words in the Icelandic Child Language Error Corpus.

Table 7 shows the top 10 most common errors in The Icelandic Child Language Error Corpus. Similar to the general corpus and the dyslexia corpus, the most common error in the child language corpus is 'punctuation', with 18.92% frequency, and the second most common error is 'wording', which comprises 10.63% of the errors in the corpus. Here we also see some predictable children's mistakes such as regarding capitalization, which is not common in the Icelandic Error Corpus, the L2 Error Corpus or the Dyslexia Error Corpus, which were written by adults.

The text collection effort lasted from February 2021 to September the same year. In order to collect the published texts, two methods were chosen; first, an open call was made to any parents with children of the appropriate age. This resulted in a few texts, but most of them were collected by means of collaborations with Icelandic elementary schools. Written assignments were collected with the help of teachers and the children's guardians signed publication agreements for publication of the texts. All authors are anonymous, have not been diagnosed with dyslexia and Icelandic is their native language.

Subcategory	Main category	Frequency (%)
punctuation	orthography	18.9
wording	style	10.6
miscellaneous	other	8.0
capitalization	orthography	6.7
insertion	vocabulary	6.1
nonword	orthography	5.9
typo	orthography	4.7
omission	vocabulary	4.6
syntax	grammar	4.3
spacing	orthography	4.2

Table 7: Top 10 most frequent errors in the Icelandic Child Language Error Corpus according to subcategory.

4 Error Analysis

The error corpora for Icelandic are of varying size (see Table 8). The proportion of L2 corpus text reflects the population of Iceland (around 15–20% of the population are immigrants), but the data from children does not reflect population numbers. Comparing error frequencies between the corpora, we can infer that people with dyslexia make the most errors and children the second most errors. However, as has been discussed previously, L2 speakers at proficiency levels A1 and A2 make more errors than both speaker groups, with 306.16 and 217.49 errors per 1,000 words, respectively.

Table 9 shows the 5 most common errors, according to subcategory, in each of the three specialized error corpora and the Icelandic Error Corpus[4] (titled 'General' here), along with each error's proportion. We see that the different error corpora share many of the most common errors, e.g. 'punctuation' and 'wording'. Note that the 'miscellaneous' subcategory only consists of dependent errors, which are often connected to errors relating

[4]This information is taken from Arnardóttir et al. (2021).

Corpus	Number of words	Number of errors	Errors per 1,000/w
L2	162,071	24,948	153.93
Dyslexia	38,891	8,436	216.91
Children	37,443	7,817	208.77

Table 8: Number of words, errors and errors per 1,000 words in the three error corpora.

General	%	L2	%	Dyslexia	%	Children	%
punctuation	25.46	wording	11.00	typo	10.7	punctuation	18.9
wording	14.74	punctuation	9.6	punctuation	10.7	wording	10.6
spacing	6.98	inflection	9.0	wording	9.6	miscellaneous	8.0
nonword	6.11	miscellaneous	7.6	nonword	9.0	capitalization	6.7
typo	5.68	agreement	6.1	miscellaneous	6.5	insertion	6.1

Table 9: Frequency of the 5 most common errors in the error corpora according to subcategory.

to wording.

The L2 error corpus is the only corpus with grammatical errors as the most common ones, i.e. 'inflection' and 'agreement', which reflects the fact that the authors' native language is not Icelandic. Inflectional errors include errors where a word is in the wrong case, e.g. when a subject of a sentence should be in the nominative case but is instead in the dative case. Agreement errors include e.g. when a finite verb is not in agreement with a noun phrase, e.g. when it comes to number.

As mentioned in Section 3.3, among the most common errors in the dyslexia corpus is 'typo', which is in accordance with what is to be expected of dyslexic writers. This error is two times more frequent in the dyslexia corpus as compared to the general corpus, and is not among the 5 most common errors in children's text. Typos are e.g. errors where a letter within a word is incorrectly replaced by a different letter, or a letter is missing within a word.

As mentioned in Section 3.4, the children's corpus shows more frequent capitalization errors than in the other corpora, but it is also the only corpus to have an error relating to vocabulary, 'insertion', among the 5 most common errors. Insertion errors are e.g. errors where a redundant conjunction or word appears.

Certain error codes only occur in certain corpora. This is illustrated in Table 10. The dyslexia corpus does not contain any unique error codes, but the L2 and children's corpora do. It is therefore possible to see if there is any specific type of error that only a certain speaker is more likely to make. Most error codes pertain to punctuation, but two errors in the L2 corpus, 'adj4noun' and 'þar4það', are both lexical. The former is when an adjective incorrectly replaces a noun and the latter is when the word *þar* 'there' is written instead of *það* 'it'.

L2	Children
adj4noun	ex4qm
þar4það	
semicolon4conjunction	
wrong-symbol	

Table 10: Error codes that only appear in certain corpora.

The error codes in each corpus were ranked by frequency of occurrence (starting with 1 for the most commonly occurring error and total number of error codes that appear in the corpus, plus 1 for the ones that never appear in it) and then compared between the corpora, using the general Icelandic Error Corpus as the default. Ranking comparison produced a *delta rank*, which is the difference between the frequency rank of a certain error between corpora, and this was extracted in Tables 11, 12, and 13, which show the 10 highest delta ranks when compared to the general corpus. This clearly shows that some error codes pertaining to grammar and lexical issues are much more frequent in the specialized corpora than in the general corpus, but interestingly enough, the delta rank is similar for each of the specialized corpora.

5 Future Use

The error corpora for Icelandic can be used separately for specific use cases, but they can also be merged for a general overview of the different types of speakers that exist in the population.

Error code	Rank General	Rank L2	Delta rank
context	132	5	127
syntax-other	132	13	113
missing-hyphen	10	104	94
v3	131	41	90
extra-sub	126	45	81
extra-prep	97	18	79
genitive	102	34	68
tense4perfect	126	63	63
extra-hyphen	46	108	62
extra-dem-pro	131	69	62

Table 11: Error codes with the highest delta rank between the general Icelandic Error corpus and the L2 corpus.

Error code	Rank General	Rank Dyslexia	Delta rank
context	132	31	101
syntax-other	132	44	88
extra-fin-verb	126	40	86
extra-sub	126	48	78
hyphen4endash	115	39	76
extra-prep	97	27	70
new-passive	124	56	68
extra-inf-part	129	63	66
v3	131	67	64
extra-dem-pro	131	69	62

Table 12: Error codes with the highest delta rank between the general Icelandic Error corpus and the Dyslexia corpus.

With the general overview, we can see how an L2 learner is going to make different errors than a native speaker, while a child will make different errors than an adult, etc. The current version of the L2 error corpus consists of texts written by adults who are only learning Icelandic. By collecting texts from learners who are learning more than one language, it would be possible to determine whether the types of errors that learners make are similar across languages, which may help in pooling larger data sources and transfer learning across languages.

Furthermore, combining all the specialized error corpora can facilitate a spellchecker that takes into account the needs of all these varieties of speakers, and can therefore detect and correct errors which are often produced by them. The error corpora can be put into practical use in creating a grammar and spelling correction software, in the same way as the Icelandic Error Corpus has been used (Óladóttir et al., 2022). Furthermore, experiments on using them as fine-tuning data for a neural spell and grammar checker have already begun (Ingólfsdóttir et al., 2022).

An error corpus can also be used to create other tools, such as language learning tools and teaching materials. Statistics and error examples from the Icelandic L2 Error Corpus can be used for developing computer-assisted language learning tools, such as flashcards (Xu and Ingason, 2021). An L2 error corpus also sheds light on learner interlanguage, which provides insight into how grammatical and lexical categories are acquired and internalized (Glisic and Ingason, 2022). This insight can be used when developing language learning tools for Icelandic, and it can also be helpful for teachers who teach Icelandic as a second language, because it helps them predict which errors the language learners will make at what stage in their proficiency level. This is also the case for teachers who teach Icelandic students with dyslexia; the Icelandic Dyslexia Error Corpus documents the most common errors made by the speakers – a valuable insight into what dyslexia looks like in Icelandic. Furthermore, the Icelandic Child Language Error Corpus can be used when teaching children how to write.

Error code	Rank General	Rank Children	Delta rank
extra-sub	126	18	108
context	132	43	89
extra-fin-verb	126	41	85
lower4upper-initial	89	8	81
extra-dem-pro	131	55	76
missing-qm	117	43	74
v3	131	59	72
syntax-other	132	61	71
extra-inf-part	129	60	69
new-passive	124	57	67

Table 13: Error codes with the highest delta rank between the general Icelandic Error corpus and the Child language corpus.

6 Conclusion

We have described three new Icelandic error corpora: the Icelandic L2 Error Corpus (Ingason et al., 2022b), the Icelandic Dyslexia Error Corpus (Ingason et al., 2022a) and the Icelandic Child Language Error Corpus (Ingason et al., 2021). All corpora are published under a CC BY 4.0 license and reflect errors made by the three user groups. The value of such corpora is diverse; they can be used to develop user-oriented spell and grammar checkers, can guide the development of language learning tools and can help in teaching Icelandic to non-native speakers, and in teaching dyslexic students or children in general to write.

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. We would like to thank the anonymous reviewers for their valuable feedback.

References

Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Karl Ingason. 2021. Creating an error corpus: Annotation and applicability. In *Proceedings of CLARIN 2021*, pages 59–63.

Isidora Glisic and Anton Karl Ingason. 2022. The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic. *Linköping Electronic Conference Proceedings*.

Anton Karl Ingason, Þórunn Arnardóttir, Lilja Björk Stefánsdóttir, and Xindan Xu. 2021. The Icelandic child language error corpus (IceCLEC) version 1.1. CLARIN-IS.

Anton Karl Ingason, Þórunn Arnardóttir, Lilja Björk Stefánsdóttir, Xindan Xu, Dagbjört Guðmundsdóttir, and Isidora Glišić. 2022a. The Icelandic dyslexia error corpus 1.2 (22.10). CLARIN-IS.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu, Isidora Glišić, and Dagbjört Guðmundsdóttir. 2022b. The Icelandic L2 error corpus (IceL2EC) 1.3 (22.10). CLARIN-IS.

Svanhvít Lilja Ingólfsdóttir, Pétur Orri Ragnarsson, Haukur Páll Jónsson, Haukur Barri Símonarson, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2022. Byte-level neural error correction model for Icelandic - yfirlestur (22.09). CLARIN-IS.

Penny Macdonald, Amparo García-Carbonell, and Sierra Jose Miguel Carot. 2013. Computer learner corpora: Analysing interlanguage errors in synchronous and asynchronous communication. *Language Learning & Technology*, 17:36–56.

Brian North and Enrica Piccardo. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion volume with new Descriptors*.

Hulda Óladóttir, Þórunn Arnardóttir, Anton Ingason, and Vilhjálmur Þorsteinsson. 2022. Developing a spell and grammar checker for Icelandic using an error corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4644–4653, Marseille, France. European Language Resources Association.

Magali Paquot and Stefan Th. Gries. 2020. *A Practical Handbook of Corpus Linguistics*. Springer Cham, Switzerland.

Xindan Xu and Anton Karl Ingason. 2021. Developing Flashcards for learning Icelandic. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 55–61, Online. LiU Electronic Press.