Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4644–4653
Marseille, 20-25 June 2022

Developing a Spell and Grammar Checker for Icelandic Using an Error Corpus

Hulda Óladóttir[1], Þórunn Arnardóttir[2], Anton K. Ingason[2], Vilhjálmur Þorsteinsson[1]
[1]Miðeind ehf., Fiskislóð 31 B/303, 101 Reykjavík, Iceland
[2]University of Iceland, Sæmundargata 2, 102 Reykjavík, Iceland
[1]{hulda, vt}@mideind.is, [2]{thar, anton}@hi.is

Abstract

A lack of datasets for spelling and grammatical error correction in Icelandic, along with language-specific issues, has caused a dearth of spell and grammar checking systems for the language. We present the first open-source spell and grammar checking tool for Icelandic, using an error corpus at all stages. This error corpus was in part created to aid in the development of the tool. The system is built with a rule-based tool stack comprising a tokenizer, a morphological tagger, and a parser. For token-level error annotation, tokenization rules, word lists, and a trigram model are used in error detection and correction. For sentence-level error annotation, we use specific error grammar rules in the parser as well as regex-like patterns to search syntax trees. The error corpus gives valuable insight into the errors typically made when Icelandic text is written, and guided each development phase in a test-driven manner. We assess the system's performance with both automatic and human evaluation, using the test set in the error corpus as a reference in the automatic evaluation. The data in the error corpus development set proved useful in various ways for error detection and correction.

Keywords: spell checker, grammar checker, spelling error correction, grammatical error correction, SEC, GEC, error corpus, Icelandic, morphologically rich languages, low-resource languages, medium-resource languages

1. Introduction

Spell and grammar checking of text is a well-established task within language technology, and checking tools are readily available for major languages. However, this task has to date not been sufficiently addressed with open and accessible solutions for Icelandic, a low-resource language (Rögnvaldsson, 2022). Spell and grammar checking was therefore included as one of the key components of the Icelandic government's strategic 5-year Language Technology Programme for Icelandic (Nikulásdóttir et al., 2020), where all deliverables are open-source and made available in the Icelandic CLARIN repository.

We describe our approach to the task, which is reproducible for other low- or medium-resource languages and especially well-suited to morphologically rich languages. The method employs different language resources, including the newly-created Icelandic Error Corpus (Arnardóttir et al., 2021), to guide the development of GreynirCorrect, a pre-existing Icelandic spell and grammar checker. The checker is available in the Icelandic CLARIN repository and on GitHub under the MIT license (Þorsteinsson et al., 2020).

The paper is structured as follows. Section 2 discusses methods for developing and evaluating spell and grammar checkers, the state of the art for Icelandic, and language-specific issues. Section 3 focuses on language resources used for the development of the Icelandic spell and grammar checker. The methods chosen for this particular project are discussed in Section 4 and the system's evaluation is described in Section 5. Future work is discussed in Section 6 and we conclude with Section 7.

2. Background

2.1. Common Methods

The spell and grammar checking task can be split into well-defined subtasks. **Error detection** involves determining whether the text, sentence or token is erroneous or deficient. **Error span detection** determines where in the text the error lies. **Error span labelling**, or error categorization, is not necessary for all uses of the tool, but is desirable if the end user is a human writing text who wants to learn from their mistakes. **Error correction** involves correcting the text, sentence, or token, and is split into spelling error correction (SEC) and grammatical error correction (GEC).

Two common approaches used in spell and grammar checking are rule-based systems on one hand and statistical or neural systems on the other, and which one is chosen depends *inter alia* on the resources at hand (Wiechetek et al., 2021).

Rule-based systems do not require as much training data as neural systems, but in turn the grammar rules, or errors, need to be listed explicitly (Deksne and Skadiņš, 2011; Fahda and Purwarianti, 2017; Jiang et al., 2012). A spell and grammar checker of this kind usually involves a parser, which analyzes a sentence and either cannot parse the sentence, meaning that it is probably erroneous, or recognizes an error pattern within it. These rules are created manually and a large error corpus is not a prerequisite for their development.

In contrast, a neural spell and grammar checker requires large amounts of text with examples of errors along with their corrections for training. Such corpora are not always readily available, particularly for low- to

medium-resource languages. The data is used to train neural networks for a correction task (Ahmadzade and Malekzadeh, 2021). One approach is training classifiers for each stage of the task and for specific error categories. Another common approach within neural spell and grammar checking is monolingual neural machine translation (Ji et al., 2017; Solyman et al., 2021; Fu et al., 2018; Jayanthi et al., 2020; Park et al., 2021). The correction of a sentence is regarded as a translation task, from the incorrect version to the correct version. In this case, only one of the three stages of spell and grammar checking is explicitly performed, i.e. error correction.

2.2. Corpora used in Spell and Grammar Checking

As mentioned, the use of error corpora when developing a spell and grammar checker is a well-established method within the field. Not only can the way in which an error corpus is used for a spell and grammar checker vary, but the error corpus itself can be of various types. For example, an error corpus can be created by manually correcting texts from informants (Flor et al., 2019; Deksne and Skadina, 2014), showing real-word spelling and grammar errors and their frequency. It can also be synthetic, where correct text is made incorrect by deliberately introducing errors (Stahlberg and Kumar, 2021). In these cases, the spelling and grammar error patterns and their frequency are determined in advance and not by their actual appearance in the training text, but these corpora are of use when time and resources are limited.

The texts included in error corpora can differ depending on their intended use, as they can be texts from native speakers (Deksne and Skadina, 2014; Rosner et al., 2012), non-native speakers (Flor et al., 2019; Volodina et al., 2016; Boyd et al., 2014; Tenfjord et al., 2006), people with dyslexia (Alamri and Teahan, 2017; Rello et al., 2014; Pedler, 2007), etc. The training domains can be kept separate for specific use cases, or merged for a more general or generic spell and grammar checker.

2.3. Icelandic-Specific Issues

Most methods are colored by the fact that the largest and most commonly available datasets are in English and focus on L2 texts, i.e. texts by second language learners. Although some Icelandic language resources exist, Icelandic is still considered a low-resource language in terms of language technology support (Rögnvaldsson, 2022). According to Rögnvaldsson (2022), Icelandic is placed on the border of the fragmentary support level and the level of weak or no support.

In contrast to English, Icelandic has a relatively free word order, which makes creating sufficient context-free grammar (CFG) rules much more difficult. Additionally, Icelandic is a morphologically rich language (MRL), which means that all inflectional information must be taken into account in the tagging analysis, not just the part-of-speech (PoS). A large portion of the word forms also has more than one possible tag and lemma. Furthermore, Icelandic is a very active compounding language, with compounds appearing as a single word with a theoretically unlimited number of constituents, requiring compound analysis for vocabulary lookup.

(1) Samsetningagreiningarvandamál
Sam-setninga-greiningar-vanda-mál
(literal meaning: *compound analysis problems*)

Errors in long compounds such as in (1) are very difficult to handle in spell checking, as the compounds are unlikely to appear in the trigrams data or in the vocabulary, even though each constituent does. Furthermore, compound analysis is challenging due to the fact that the semantic relationship between the parts of a compound is subject to nuanced variation (Ingason and Sigurðsson, 2020).

Error correction in Icelandic is also complicated by morphosyntactic variation and ongoing language change – and prescriptive efforts to standardize usage in the domains in question. While some error categories are widely attested across languages, some such patterns of variation are language-specific. The most well-known case is the so-called *Dative Substitution*, exemplified by (2) and (3), and extensively documented in the literature (Jónsson and Eythórsson, 2005; Ingason, 2010; Thráinsson, 2013; Jónsson, 2013; Þráinsson et al., 2015; Nowenstein and Ingason, 2021).

(2) Mig langar í jarðarber.
me.ACC longs in strawberries.
'I want strawberries.'

(3) Mér langar í jarðarber.
me.DAT longs in strawberries.
'I want strawberries.'

Dative Substitution affects verbs with experiencer subjects, with dative case (3) replacing the accusative case (2) that used to be the subject case for these verbs. While nominative is the default subject case in Icelandic, here we have a context where two non-default cases compete for use in the language community, and a prescriptive demand to point users toward the accusative in error correction.

The so-called *New Passive* (or *New Construction; New Impersonal*) is another widely discussed case of variation in Icelandic morphosyntax (Maling and Sigurjónsdóttir, 2002; Eythórsson, 2008; Jónsson, 2009; Sigurðsson, 2011; Ingason et al., 2012; Sigurðsson, 2017). Consider the active in (4).

(4) Álfurinn lamdi strákinn. (Active)
the.elf.SBJ beat the.boy.OBJ
'The elf beat the boy.'

This active sentence corresponds to the *Canonical Passive* in (5), in which the object of the active is realized

with subject properties. However, in the innovative New Passive in (6), a passive participle is used while the object of the active remains an object.

(5) Strákurinn var laminn. (Canonical P.)
the.boy.SBJ was beaten.PASS
'The boy was beaten.'

(6) Það var lamið strákinn. (New P.)
there was beaten.PASS the.boy.OBJ
'The boy was beaten.'

While the New Passive has been gaining ground in the language community, it has not been accepted into the prescriptive standard, and thus it is feasible that an error detection system points a user towards the Canonical Passive.

2.4. Spell and Grammar Checking for Icelandic

Spell and grammar checking for Icelandic is largely lacking, particularly in open-source and readily available solutions, and most tools only tackle spell checking for context-independent spelling errors. Below is an overview of attempts at spell and grammar checking and available resources.

Hunspell,[1] a spell checker for various text editors, has been implemented for Icelandic. It includes a thesaurus and a dictionary with tags and paradigms. It uses n-grams, a vocabulary and rule-based pronunciation data.

Aspell[2] uses the metaphone algorithm to find possible suggestions for spelling errors.

Púki[3] is closed, commercial spell-checking software that is integrated into common text editors once the user has obtained it. It includes a thesaurus and can therefore suggest synonyms for words in a text and learn new words and terms from the text itself. Its vocabulary consists of word stems and affixation rules.

Stafsetning 2004/2010 was a closed-source spell checker for Mac OS but is unavailable now.

Microsoft Editor[4] offers spell and grammar checking, along with analysis of readability, conciseness and other refinements. Only spell checking for context-independent errors and a hyphenation tool have been implemented for Icelandic.

Some spell checkers can also handle context-sensitive spelling errors, using confusion sets and other similar methods.

A **research project** used Naive Bayes and Winnow classifiers along with selected confusion sets to detect and correct context-sensitive spelling errors (Ingason et al., 2009). In a similar vein, an extensive collection of confusion sets along with their respective frequencies has been prepared (Friðriksdóttir and Ingason, 2020).

LanguageTool[5] uses XML rules to detect and correct spelling errors. Rules for Icelandic spelling errors have been defined but are not included in the current version.

The **Skrambi** system is only available through a limited online user interface.[6] It was originally developed for converting typed, handwritten or printed text into machine-encoded text, i.e. for optical character recognition (Daðason, 2012). The errors which arose in the conversion were corrected using the software. It was then further developed using confusion sets and is capable of context-sensitive spellchecking.

GreynirCorrect[7] is the only open-source Icelandic spell and grammar checker, and is the spell and grammar checker whose development is described in this paper. It handles context-independent and context-sensitive spelling errors like the aforementioned tools, as well as grammar errors and select style errors.

2.5. Evaluation Methods

The most common metrics when evaluating spell and grammar checkers are the following:

- **Token-based detection**: An error is detected within the correct token.
- **Span-based detection**: An error is detected within the correct span.
- **Span-based correction**: An error is detected and corrected within the correct span.

Table 1 gives examples of how these three metrics function. The original, incorrect sentence is *Ár var ný* 'A year was new', with the span 2–3 and the corrected word *nýtt* 'new'. These are shown as [2, 3, nýtt]. The bottom three rows in the leftmost column represent possible hypotheses that the system could produce, and the corresponding results for each metric are shown in the following columns.

Ár var ný [2, 3, nýtt]	Span-based correction	Span-based detection	Token-based detection
[2, 3, nýtt]	Yes	Yes	Yes
[2, 3, nýr]	No	Yes	Yes
[1, 2, nýtt]	No	No	Yes

Table 1: An example of each evaluation metric.

Performance is evaluated with an $F_{0.5}$ score, which gives precision twice the weight of recall. This has stronger correlation with human ratings and provides a better user experience, as it is more important to users to avoid incorrect annotations than to find all the errors in the text. GLEU (Napoles et al., 2015) is another

[1] https://github.com/nifgraup/hunspell-is
[2] http://aspell.net/
[3] https://www.puki.is/
[4] https://www.office.com/
[5] https://languagetool.org/
[6] http://skrambi.arnastofnun.is
[7] https://github.com/mideind/GreynirCorrect

metric available for spell and grammar checking and is a variant of the BLEU metric, used in machine translation. GLEU uses weighted precision of n-grams over the reference.

The most widely-used datasets for grammatical error correction (GEC) for English are the CoNLL-2014 shared task test set[8] (Ng et al., 2014) and the BEA shared task – 2019 dataset[9] (Bryant et al., 2019). These have been manually annotated with corrections and error categories.

A different viewpoint on the performance of a spell and grammar checker can be established with the *closest-gold* metric (Rozovskaya and Roth, 2021). The datasets above mostly follow the principle of 'minimal edits', i.e. carrying out as few edits as possible that result in a valid sentence. The same principle is followed in the closest-gold metric, except that the starting point is not the original text, but the system output. This provides a much fairer reference, as there are in many cases more than one possible way of correcting an error. According to our experiments, this results in a performance 10–25 percentage points better than standard evaluations.

3. Language Resources

Several language resources are used for developing the spell and grammar checker, some of which were created with that purpose in mind. Their role in the development is described further in Section 4.3.

3.1. The Icelandic Error Corpus

In order to improve the spell and grammar checker, an Icelandic error corpus was created: the Icelandic Error Corpus (IceEC; Arnardóttir et al., 2021; Ingason et al., 2021c). This is a collection of texts in modern Icelandic written by native speakers, consisting of 56,794 errors. These texts have been manually annotated for errors and therefore reflect real-word spelling and grammar errors made by Icelandic informants. The corpus is split into two parts in order to use it for guiding the development of the spell and grammar checker, i.e. a training set and a test set. The training set comprises 90% of the corpus, 52,312 errors, while the test set comprises 10%, 4,482 errors.

The annotation scheme used in the Icelandic Error Corpus was specifically created to reflect the errors in the corpus instead of adhering to an existing annotation scheme, and to aid in improving the spell and grammar checker. The annotation scheme consists of three hierarchical levels: main categories, subcategories and error codes. Each main category consists of several subcategories. Each subcategory in turn consists of several error codes, and errors in the corpus are annotated with error codes. The main categories and subcategories are used to guide the development of the spell and grammar checker, using the training set, while improving the checker and the test set to measure the improvement.

The annotation scheme consists of 6 main categories, 32 subcategories and 258 error codes. The main categories are:

- Coherence: errors related to grammatical context within a text.
- Grammar: errors related to agreement, aspect, case, syntax and such.
- Orthography: errors that only affect a word's appearance, e.g. spelling and capitalization.
- Other: a dependent error.
- Style: style errors such as using foreign words, symbols or a particular register.
- Vocabulary: semantic errors or deficiencies.

3.2. Icelandic Standards

In addition to the Icelandic Error Corpus, a few grammar and spelling standards for Icelandic were used. The Icelandic language council's spelling rules[10] were used to determine corrected values for orthographic errors and to determine rules on punctuation, for example which hyphens should be used under which circumstances. In order to correct various errors relating to language usage, a resource called *Málfarsbankinn*[11] (direct translation: *The Language Usage Bank*) was used. This is a collection of rules and general advice concerning grammar, fixed phrases, spelling, and more.

These standards also guided proofreading when the Icelandic Error Corpus was created, and were used to determine whether an error was present and how it should be corrected.

3.3. Miscellaneous Language Resources

Among other large language resources we used were the DIM, the IGC and a trigram model. Additionally, we used many smaller wordlists detailed in the development section. The **Database of Icelandic Morphology (DIM)**[12] is a multipurpose database that, amongst other things, contains approx. 287,000 lemmas and their inflectional paradigms (Bjarnadóttir et al., 2019). To enable fast lookup, DIM has been compressed and encapsulated in a Python package (Þorsteinsson et al., 2021a). A recent version of DIM added *Ritmyndir*, a collection of token-level errors linked to correct values, tags, error categorization, and the appropriate part of The Icelandic Standards, where available.

The **Icelandic Gigaword Corpus**[13] (IGC) is a large corpus of approx. 1550 million running words of text

[8] https://www.comp.nus.edu.sg/~nlp/conll14st.html

[9] https://www.cl.cam.ac.uk/research/nl/bea2019st/

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

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

[12] https://bin.arnastofnun.is/DMII/

[13] http://igc.arnastofnun.is/

(Steingrímsson et al., 2018). Texts come from varied sources, such as news media, books, parliamentary texts, and social media. Each word is attached to a morphosyntactic tag and lemma.

The IGC was used to train a **trigram model** (Þorsteinsson and Óladóttir, 2020) on a random sample of curated document collections from the corpus. Duplications and texts from before 1980 were removed, along with texts from sources deemed likely to have below-average proofreading standards. The result was over 100 million tokens from the corpus. Token-level correction was carried out to ensure the trigrams contained only correct word forms. In the first version of the trigram dataset, we ran into cases where certain very common errors occurred more frequently in the text than the correct forms, resulting in the correct version being annotated as an error and the error being suggested. Trigrams with a frequency of 1 or 2 were also removed, as they were more likely to contain errors. The final trigrams dataset includes over 14 million unique trigrams as well as frequency data for unigrams, bigrams and trigrams.

4. Methods

The system is built with a rule-based tool stack consisting of a tokenizer, a morphological tagger, and a parser, described in (Þorsteinsson et al., 2019). Token-level errors are generally dealt with at the tokenization and tagging stages, and sentence-level errors during or after the parsing stage, by searching the syntactic tree. This process is depicted in Figure 1 and described in the subsections below.

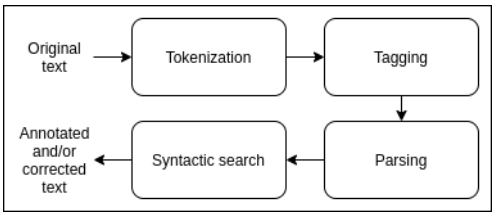

Figure 1: A diagram of the system flow.

4.1. Token-Level Error Annotation

The tokenizer (Þorsteinsson et al., 2021c) is organized in layers, with each additional layer recognizing more complex tokens. In the process, errors in punctuation are detected and corrected/normalized. These include foreign quotation marks[14] or double commas instead of quotation marks, wrong dashes for the context, three periods instead of an ellipsis, multiple punctuation marks (',.' is a very common error), and punctuation in abbreviations. The original text for each token is preserved, so the spell checker module can add error annotations with an intact reference to the original.

[14]Icelandic uses German-style double quotes, i.e. „these".

The output of the basic tokenization process, i.e. splitting text into sentences and tokens, is then sent through an error-detection layer designed to detect context-independent token-level errors. These include duplicated words, single-word compounds that have erroneously been split into two or more tokens, and phrases that have been written as a single amalgamated word. These errors are in most cases found via lookups in lists of words and morphemes from configuration files.

Icelandic words can be highly ambiguous, so information on part-of-speech, lemma, and inflectional attributes is necessary. The tagger, also built in layers, uses lookup in DIM and compound analysis to collect all possible tags for each word token. A final tag is not selected at this stage; that task is left to the parser. With this information, we can handle more complex token-level errors and deficiencies (such as more complex splitting errors not reliant on word lists), capitalization errors, and taboo words.

Common context-dependent errors that only rely on context from the nearest neighbors are handled with a list of semi-fixed phrases and common erroneous variations. The list contains lemmatized forms of all words, in order to facilitate matching with inflected word forms. Fixed phrases are handled in the same way without lemmatization.

For unknown or rare words, we collect all possible substitutes found either in IGC, DIM or auxiliary vocabularies with a Levenshtein distance of 1. The trigram language model discussed in Section 3.3 is then used to rank the substitutes by probability given the context. If a substitute is deemed probable enough, the original word is annotated as an error and the substitute is given as a possible correction. The tags for the possible substitutes are added to the set of possible tags for the tokens.

4.2. Sentence-Level Error Annotation

The underlying parser uses context-free grammar rules (CFG rules) to form all possible valid syntactic structures (a parse forest) for each sentence. Instead of a pipeline of tokenizing, tagging (in the sense of selecting a single most probable tag for each word) and then parsing, the latter two are done at the same time. Due to this, the tagger and parser (Þorsteinsson et al., 2021b) are bundled together.

All possible tags, both for the original word forms and potential corrections, are collected in the tagging layer and sent to the parsing stage. The parser then selects the tags that can fit into valid syntactic structures, in a whole-sentence context. All such possible sentence trees are collected and heuristically ranked from the bottom up, with the top-scoring family of children being selected at each parent nonterminal node. This method gives us a much wider context for the tagging task, and does not pigeonhole the parser into using tags that cannot appear in a valid parse tree for the entire sentence. Since we also added tags for all possible to-

ken corrections, we can now make a much better informed decision on whether the token contains an error or not, using the syntactic context.

Specific erroneous grammar rules were added to the parser to handle well-known invalid structures, such as the Dative Substitution described in Section 2.3. These rules are not included in standard parsing, except for a few very common error rules, and are only available to the parser when it is used for grammar error checking. The error rules describe well-defined errors pertaining to syntax, resulting in invalid sentences. Some erroneous syntactic nodes can be directly mapped to correct ones, but in other cases they are searched for after parsing and corrected at that stage.

As an example, in Figure 2, an instance of Dative Substitution results in a grammar error being detected by an error grammar rule and annotated, despite the relative clause appearing between the head noun of the subject and the main verb. This error is then readily correctable by casting the subject noun phrase in the parse tree from the dative to the accusative case.

Figure 2: An example of a grammar error, as displayed in the GreynirCorrect web user interface. The sentence is *The man who fell into the pond needs a hammer*. The verb *vanta* (need) takes a subject in the accusative (*Manninn*) instead of the erroneous dative (*Manninum*). The text boxes explain the error.

Finally, we search for questionable syntactic patterns in the parse tree for each sentence to find grammar errors. These errors include wording errors, such as attaching the wrong prepositional phrase to a verb or giving an object the wrong case, resulting in a sentence that is strictly speaking syntactically valid, but morphologically or semantically invalid. The search function implements regex-like patterns for syntactic trees, with patterns given as strings. The patterns allow matching for literal text, specific terminals and non-terminals, and wildcards. It also allows sequential and hierarchical matching. Each error rule and search pattern is handled in its own function. There, the error span is further refined if necessary, a correction provided if possible, and details collected about what the error entails by probing the parse tree. Lastly, we attach an error annotation to the span containing this information.

4.3. Development

We used test-driven development to ensure the best results. The Icelandic Error Corpus development set provided frequency information for each error category, and the test set was used for automatic evaluation, giving an $F_{0.5}$ measure for each error category. This information was used to guide the development iteratively; the most frequent categories with the lowest scores were prioritized at each stage.

The development set was also useful for looking at examples within a single error category. Common cases in the category were detected and handled. We reviewed how the checker handled the examples and subsequently discovered false negatives and bugs in the handling. Examples for each type of error were added to the package's test suite, so regressions could be detected as early as possible.

The spell and grammar checker's development was also data-driven. For token-level annotation, we used DIM as our valid vocabulary, along with compound analysis. As error data, we used a list of non-words from the development set in IceEC (Arnardóttir and Ingason, 2020a). Additionally, we used a list of non-words from earlier lexical acquisition, common and systematic inflectional errors, systematic spelling errors (Arnardóttir and Ingason, 2020b) and a list of search queries on the DIM website that did not match any entry (Arnardóttir et al., 2020). All these word lists contain the error, the correct value and in some cases the correct PoS tag.

For sentence-level annotation, we relied mostly on guidance from The Icelandic Standards, and the development set of IceEC.

5. Evaluation

In order to evaluate the spell and grammar checker, three methods are utilized: automatic evaluation using manually annotated data, the closest-gold metric and human evaluation. These three methods use different resources to deliver results on different aspects of the spell and grammar checker, and together, they report on the overall performance of the spell and grammar checker. We use $F_{0.5}$ measures instead of F_1, which weigh precision higher than recall, as discussed in Section 2.5. This is in alignment with our task, as it is more important to accurately report and correct an error than it is to report and correct all of them, i.e. we want to minimize false positives.

Our objective is for the system to help rather than hinder text writing in general. State-of-the-art systems for English, which, as discussed in Section 2.3, has different challenges to Icelandic, use neural models to reach $F_{0.5}$ measures of 65–76 (Tarnavskyi, 2021; Rothe et al., 2021). These results provide some insight into where we want to head with our currently rule-based system.

5.1. Automatic Evaluation

To date, the focus has been on error detection out of the subtasks discussed in 2.1, to ensure the best coverage

of error categories before delving deeper. We started with token-level errors and then tackled sentence-level grammar and vocabulary errors.

Using span-based detection, the system reaches an $F_{0.5}$ measure of 73.41 for token-level errors, excluding punctuation errors. Typos, a third of the token-level errors in the test data, reach 92.66. The system currently reaches an $F_{0.5}$ measure of 29.35 for sentence-level grammar and vocabulary errors. Table 2 displays the precision, recall and $F_{0.5}$ measure for each subcategory within *orthography*, *grammar* and *vocabulary* in the test set, along with their frequency in said set. These results should be viewed through the lens of the language-specific issues discussed in Section 2.4.

Comparing these results to the accuracy of other systems available for Icelandic is difficult for several reasons. First, the other systems only handle spell checking, and mostly at the token-level, so only a part of the system discussed in the paper could be compared to the other systems. Second, the systems and the corpora do not use the same error schema, and third, the errors do not exist in a vacuum, so false positives for grammar errors, as an example, can obfuscate results for token-level errors.

Subcategory	Prec.	Rec.	$F_{0.5}$	Freq.
Orthography				
Punctuation	84.24	36.95	47.56	498
Typo	100.00	72.86	92.66	210
Spacing	96.89	55.50	77.62	209
Capitalization	63.26	21.93	44.86	114
Nonword	48.99	74.32	44.84	74
Spelling	87.50	65.50	80.95	60
Grammar				
Agreement	45.31	15.79	32.15	76
Prep	90.22	13.33	22.48	45
Mood	3.43	22.22	4.13	27
Inflection	100.00	7.69	29.41	13
Syntax	38.46	15.38	29.59	13
Aspect	4.00	25.00	4.81	4
Case	100.00	25.00	62.50	4
Vocabulary				
Insertion	90.00	33.33	52.91	18
Collocation	73.33	26.67	54.32	15
Semantic	58.21	14.29	25.10	14

Table 2: Precision, recall, $F_{0.5}$ measure for span-based detection and frequency for each orthography, grammar and vocabulary subcategory found in the test set. The subcategory 'prep' is short for 'preposition'.

5.2. Closest-Gold Metric

To acquire a more comprehensive understanding of performance, the system output for 100 sentences from the development set was manually annotated according to the closest-gold metric discussed in Section 2.5. Table 3 displays the closest-gold results of token-level errors, i.e. orthographic errors, and sentence-level errors, i.e. grammatical errors and errors relating to vocabulary, compared to the original results. Although the sample is small, the comparison suggests that the system performs better than using the IceEC as a reference reflects.

Category	**IceEC**	**CG**
Token-level	73.41	86.48
Sentence-level	29.35	51.47

Table 3: Comparison of $F_{0.5}$ measures for IceEC references and closest-gold references.

5.3. Human Evaluation

To obtain a better picture of the user experience, the system was integrated into the editorial environment of an online news media company to carry out user tests. Journalists at the media company in question used the spell and grammar checker when writing news articles and other items, and provided feedback. The feedback received roughly corresponds to the error detection and error correction metrics, keeping track of whether the user accepted or rejected each proposed correction. If the correction was accepted, all metrics in Table 1 were positive. If the correction was rejected, resulting in a false span-based correction, the user detailed why. Was the original text correct (false positive for error detection), was the original text correct but the correction wrong (true positive for error detection, false negative for error correction), or was there some other reason? This information was sent automatically to a database, along with information on the error code, word span of the error, the original text of the span, the corrected version of the span, and the complete sentence. In addition to this, the journalists provided informal, verbal feedback. The feedback was used to improve the user experience, with bug fixes and implementation of several options for different use cases, such as getting a list of the five most likely suggested corrections for the user to select from. The overall feedback indicated that the checker was a beneficial addition to the editorial workflow.

During development, a regularly updated version of the spell and grammar checker has been accessible to the public via a website, and valuable feedback has been received from users.

6. Future Work

The focus of the next development phase will on be providing better corrections, categorization, details and guidance for the end user. New data pertaining to specific error types and/or the Icelandic Standards will be incorporated, such as confusion sets (Friðriksdóttir and Ingason, 2020) and relevant links to the Icelandic Standards. We will also experiment with neural methods, such as a binary classifier for determining whether a

sentence is likely to contain an error, classifiers for specific error categories which are difficult to handle with rule-based methods, and a translation model to 'translate' incorrect text into correct Icelandic. This work includes automatically generating erroneous text with an error distribution similar to the expected distribution, which is then used as training data for the translation model. The spell and grammar checking as a whole will move towards the sentence-level, both in checking and evaluation, to assess more obscure issues such as style, fluency, etc. Lastly, the system will be made more accessible via integration into common text editors and editorial environments, such as those employed in the user evaluation discussed in Section 5.3.

The methods described above for improving a spell and grammar checker for native Icelandic speakers can also be used to develop a specific spell and grammar checker for non-native speakers, people with dyslexia or children. Error corpora have been created for these respective informant groups, similar to the Icelandic Error Corpus. The Icelandic L2 Error Corpus (Ingason et al., 2021d) consists of texts written by second-language users of Icelandic, the Icelandic Dyslexia Error Corpus (Ingason et al., 2021b) consists of texts written by native Icelandic speakers with dyslexia, and the Icelandic Child Language Error Corpus (Ingason et al., 2021a) consists of texts written by children aged 10–15. These error corpora have the same annotation scheme and structure as the Icelandic Error Corpus and can therefore be used in the same way to create spell and grammar checkers specifically for these groups.

7. Conclusion

In this paper, we have presented a spell and grammar checker for Icelandic that uses an error corpus at all stages of development, along with other available resources. The system is built with a rule-based tool stack, handling both token-level and sentence-level error annotation. We assess the system, both with automatic evaluation and human evaluation, using $F_{0.5}$ for error span detection. The results indicate that the methods described are viable for creating a spell and grammar checker for Icelandic, but the methodology is also applicable when developing spell and grammar checkers for other morphologically rich and/or low- to medium-resource languages.

The spell and grammar checker is the first open-source system to tackle grammar checking for Icelandic, and is published under the MIT license in the Icelandic CLARIN repository (Þorsteinsson et al., 2020).

8. 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 Kristín Bjarnadóttir, editor of DIM, for her valuable feedback and data, and our fellow members of the Consortium on Icelandic Language Technology. We would also like to thank the anonymous reviewers for their valuable feedback.

9. Bibliographical References

Ahmadzade, A. and Malekzadeh, S. (2021). Spell correction for Azerbaijani language using deep neural networks.

Alamri, M. and Teahan, W. J. (2017). A new error annotation for dyslexic texts in Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 72–78, Valencia, Spain, April. Association for Computational Linguistics.

Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B., and Ingason, A. K. (2021). Creating an error corpus: Annotation and applicability. In *Proceedings of CLARIN 2021*, pages 59–63, September.

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The database of Icelandic morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NODALIDA 2019)*, pages 146–154, Turku, Finland.

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).

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, August. Association for Computational Linguistics.

Daðason, J. (2012). Post-correction of Icelandic OCR text. Master's thesis, University of Iceland.

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.

Deksne, D. and Skadiņš, R. (2011). CFG based grammar checker for Latvian. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 275–278, Riga, Latvia, May. Northern European Association for Language Technology (NEALT).

Eythórsson, T. (2008). The New Passive in Icelandic really is a passive. In Thórhallur Eythórsson, editor, *Grammatical Change and Linguistic Theory: The Rosendal Papers*, pages 173–219. John Benjamins, Amsterdam.

Fahda, A. and Purwarianti, A. (2017). A statistical and rule-based spelling and grammar checker for Indonesian text. In *2017 International Conference on Data and Software Engineering (ICoDSE)*, pages 1–6.

Flor, M., Fried, M., and Rozovskaya, A. (2019). A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 76–86, Florence, Italy, August. Association for Computational Linguistics.

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)*.

Fu, K., Huang, J., and Duan, Y. (2018). Youdao's winning solution to the NLPCC-2018 task 2 challenge: A neural machine translation approach to Chinese grammatical error correction. In Min Zhang, et al., editors, *Natural Language Processing and Chinese Computing*, pages 341–350, Cham. Springer International Publishing.

Ingason, A. K. and Sigurðsson, E. F. (2020). Attributive compounds. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

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., Legate, J. A., and Yang, C. (2012). The evolutionary trajectory of the Icelandic new passive. *University of Pennsylvania Working Papers in Linguistics*, 19(2):11.

Ingason, A. K. (2010). Productivity of non-default case. *Working papers in Scandinavian syntax*, 85:65–117.

Jayanthi, S. M., Pruthi, D., and Neubig, G. (2020). NeuSpell: A neural spelling correction toolkit.

Ji, J., Wang, Q., Toutanova, K., Gong, Y., Truong, S., and Gao, J. (2017). A nested attention neural hybrid model for grammatical error correction.

Jiang, Y., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X., Wang, C., and Zhang, W. (2012). A rule based Chinese spelling and grammar detection system utility. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 437–440.

Jónsson, J. G. (2009). The new impersonal as a true passive. In Artemis Alexiadou, et al., editors, *Advances in comparative Germanic syntax*, pages 281–306. John Benjamins, Amsterdam.

Jónsson, J. G. and Eythórsson, T. (2005). Variation in subject case marking in Insular Scandinavian. *Nordic Journal of Linguistics*, 28.2:223–245.

Jónsson, J. G. (2013). Two types of case variation. *Nordic Journal of Linguistics*, 1(36):5–25.

Maling, J. and Sigurjónsdóttir, S. (2002). The new impersonal construction in Icelandic. *The Journal of Comparative Germanic Linguistics*, 5(1):97–142.

Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China, July. Association for Computational Linguistics.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June. Association for Computational Linguistics.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France, May. European Language Resources Association.

Nowenstein, I. and Ingason, A. K. (2021). Featural dynamics in morphosyntactic change. In Jóhannes G. Jónsson et al., editors, *Syntactic features and the limits of syntactic change*. Oxford University Press, Oxford.

Park, C., Kim, K., Yang, Y., Kang, M., and Lim, H. (2021). Neural spelling correction: translating incorrect sentences to correct sentences for multimedia. *Multimedia Tools and Applications*, 80:34591–34608, November.

Pedler, J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. Ph.D. thesis, Birkbeck, London University.

Þráinsson, H., Eyþórsson, Þ., Svavarsdóttir, Á., and Blöndal, Þ. (2015). "Fallmörkun." ['Case marking.']. In Höskuldur Þráinsson, et al., editors, *Tilbrigði í íslenskri setningagerð II*. Málvísindastofnun Háskóla Íslands, Reykjavík.

Rello, L., Baeza-Yates, R., and Llisterri, J. (2014). DysList: An annotated resource of dyslexic errors. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1289–1296, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Rögnvaldsson, E. (2022). Report on the Icelandic language.

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 Re-*

sources and Evaluation (LREC'12), pages 743–750, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rothe, S., Mallinson, J., Malmi, E., Krause, S., and Severyn, A. (2021). A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830*.

Rozovskaya, A. and Roth, D. (2021). How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698, Online, April. Association for Computational Linguistics.

Sigurðsson, E. F. (2017). *Deriving case, agreement and Voice phenomena in syntax*. Ph.D. thesis, University of Pennsylvania.

Sigurðsson, H. Á. (2011). On the new passive. *Syntax*, 14(2):148–178.

Solyman, A., Zhenyu, W., Qian, T., Elhag, A. A. M., Toseef, M., and Aleibeid, Z. (2021). Synthetic data with neural machine translation for automatic correction in Arabic grammar. *Egyptian Informatics Journal*, 22(3):303–315.

Stahlberg, F. and Kumar, S. (2021). Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online, April. Association for Computational Linguistics.

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.

Tarnavskyi, M. (2021). Improving sequence tagging for grammatical error correction. Master's thesis, Ukrainian Catholic University.

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).

Thráinsson, H. (2013). Ideal speakers and other speakers. The case of dative and other cases. In Beatriz Fernández et al., editors, *Variation in Datives – A Micro-Comparative Perspective*, pages 161–188. Oxford University Press.

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.

Wiechetek, L., Pirinen, F., Hämäläinen, M., and Argese, C. (2021). Rules ruling neural networks – neural vs. rule-based grammar checking for a low resource language. In *Proceedings of the International Conference Recent Advances In Natural Language Processing 2021*, International conference Recent advances in natural language processing, pages 1530–1539. INCOMA.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. (2019). A wide-coverage context-free grammar for Icelandic and an accompanying parsing system. In *Proceedings of the International Conference on Recent Advances in Natural Language Proceedings (RANLP 2019)*, pages 1397–1404, Varna, Bulgaria.

10. Language Resource References

Arnardóttir, Þ. and Ingason, A. K. (2020a). Icelandic error corpus nonwords. CLARIN-IS.

Arnardóttir, Þ. and Ingason, A. K. (2020b). nonwords. CLARIN-IS.

Arnardóttir, Þ., Andrésdóttir, Þ. D., Árnason, Þ. A., Hafsteinsdóttir, H., Þórisson, S., Sigurðsson, E. F., and Bjarnadóttir, K. (2020). Icelandic search query errors (IceSQuEr) 0.1. CLARIN-IS.

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. (2021a). The Icelandic Child Language Error Corpus (IceCLEC) version 1.1. CLARIN-IS.

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. (2021b). The Icelandic Dyslexia Error Corpus (IceDEC) version 1.1. CLARIN-IS.

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

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I. (2021d). The Icelandic L2 Error Corpus (IceL2EC) version 1.2. CLARIN-IS.

Þorsteinsson, V. and Óladóttir, H. (2020). Icegrams (2020-09-30). CLARIN-IS.

Þorsteinsson, V., Óladóttir, H., Arnardóttir, Þ., and Þórðarson, S. (2020). GreynirCorrect (2021-09-30). CLARIN-IS.

Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. (2021a). BinPackage. CLARIN-IS.

Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. (2021b). GreynirPackage. CLARIN-IS.

Þorsteinsson, V., Óladóttir, H., Þórðarson, S., and Ragnarsson, P. O. (2021c). Tokenizer. CLARIN-IS.