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

Hulda Óladóttir\textsuperscript{a}, Pórunn Arnardóttir\textsuperscript{b}, Anton Karl Ingason\textsuperscript{b}, Vilhjálmur Porsteinsson\textsuperscript{a}

\textsuperscript{a}Mi\=wind ehf., Fiskiós\=óð 31 B/303, 101 Reykjavík, Iceland, \textsuperscript{b}University of Iceland, Sæmundargata 2, 102 Reykjavík, Iceland

Introduction
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 GreyhirCorrect, the first open-source spell and grammar checker tool for Icelandic, using the newly-created Icelandic Error Corpus at all stages. The project was funded as one of the key components of the Icelandic government’s strategic 5-year Language Technology Programme for Icelandic.

Language Resources
Several language resources are used to guide the development of the spell and grammar checker, providing information on spelling rules, language usage, lemmas, inflectional paradigms, morphosyntactic tags and trigrams. The resources include:
- The Icelandic language council’s spelling rules and the Language Usage Bank.
- The Database of Icelandic Morphology (DIM)
- The Icelandic Gigaword Corpus
- An Icelandic trigram language model

The Icelandic Error Corpus
A dataset, The Icelandic Error Corpus, was created in order to guide the development of the spell and grammar checker and to measure improvements.
- The corpus is a collection of real-word spelling and grammar errors made by Icelandic informants.
- Consists of roughly 60,000 errors in manually corrected texts.
- Errors are categorized according to an annotation scheme, which consists of three hierarchical levels: main categories, subcategories and error codes.
- Split up into a development (90%) and test set (10%).
  - The development set provides frequency information on error categories and is used to develop the spell and grammar checker.
  - The test set is used for automatic evaluation, giving an $F_{0.5}$ measure for each error category, thereby measuring improvements in the spell and grammar checker.

The Spelling and Grammar Checker
The GreyhirCorrect system is built with a rule-based tool stack consisting of a tokenizer, a morphological tagger, and a parser. The system is roughly split into token-level error annotation and sentence-level error annotation.

Token-level error annotation
- Errors in punctuation are detected and corrected/normalized in the tokenizer.
- Context-independent token-level errors are detected after the basic tokenization, such as duplicated words, and word splitting.
- Tag information is necessary for capitalization errors, taboo words, and more complex splitting errors.
- Semi-fixed phrases along with common erroneous variations (allowing for inflection) handle some common context-dependent errors.
- For unknown or rare words, all possible substitutes with a Levenshtein distance of 1 are collected and ranked with a trigram language model.

Sentence-level error annotation
- Specific erroneous grammar rules in the underlying parser recognize well-known invalid syntactic structures, such as Dative Substitution.
- Questionable syntactic patterns in the parse tree for each sentence are used to detect grammar errors, such as attaching the wrong prepositional phrase to a verb or giving an object the wrong case.

Evaluation

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>Prec.</th>
<th>Rec.</th>
<th>$F_{0.5}$</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthography</td>
<td>85.22</td>
<td>49.1</td>
<td>62.37</td>
<td>1165</td>
</tr>
<tr>
<td>Grammar</td>
<td>53.91</td>
<td>15.93</td>
<td>25.29</td>
<td>182</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>75.21</td>
<td>25.53</td>
<td>45.08</td>
<td>47</td>
</tr>
</tbody>
</table>

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, and feedback that roughly corresponds to the error detection and error correction metrics was collected.

Conclusion
The results indicate that our methods are viable for creating a spell and grammar checker for Icelandic and 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 and on GitHub.