Disambiguating Confusion Sets as an Aid for Dyslexic Spelling

Steinunn Rut Friðriksdóttir, Anton Karl Ingason
Faculty of Icelandic and Comparative Cultural Studies
University of Iceland, Sæmundargata 2, 102 Reykjavík
srf2, antoni@hi.is

Abstract
Spell checkers and other proofreading software are crucial tools for people with dyslexia and other reading disabilities. Most spell checkers automatically detect spelling mistakes by looking up individual words and seeing if they exist in the vocabulary. However, one of the biggest challenges of automatic spelling correction is how to deal with real-word errors, i.e. spelling mistakes which lead to a real but unintended word, such as when *then* is written in place of *than*. These errors account for 20% of all spelling mistakes made by people with dyslexia. As both words exist in the vocabulary, a simple dictionary lookup will not detect the mistake. The only way to disambiguate which word was actually intended is to look at the context in which the word appears. This problem is particularly apparent in languages with rich morphology where there is often very little orthographic difference between grammatical items. In this paper, we present our novel confusion set corpus for Icelandic and discuss how it could be used for context-sensitive spelling correction. We have collected word pairs from seven different categories, chosen for their homophonous properties, along with sentence examples and frequency information from said pairs. We present a small-scale machine learning experiment using a decision tree binary classification which results range from 73% to 86% average accuracy with 10-fold cross validation. While not intended as a finalized result, the method shows potential and will be improved in future research.

Keywords: homophones, dyslexia, reading disabilities, confusion sets, disambiguation, context dependency, Icelandic

1. Introduction
According to Mody and Silliman, dyslexia accounts for 80% of diagnosed learning disabilities. It causes problems with the mapping process between orthographic and phonological words and parts (Mody and Silliman, 2008). This means that dyslexic individuals might show difficulties in word segmenting as well as phoneme identification and manipulation. There are two main types of orthographic errors considered when evaluating effects of dyslexia on spelling. Phonetically accurate errors include, for example, adding an unnecessary double consonant or omitting a silent letter, resulting in plausible orthographic representations of the phonemes in question. Phonetically inaccurate errors include phoneme omissions, additions and substitutions which cannot be taken to represent the phonemes in the intended word (Bernstein, 2009).

One possible representation of phonetically accurate spelling mistakes is the substitution of homophones. Examples of this include when *then* is written in place of *than* or when *by* is written in place of *buy*. Since these mix-ups result in unintended but valid words, they often go undetected by spell checkers and other proofreading software which would otherwise pick up on an out-of-vocabulary spelling mistake. Bernstein also notes in his paper that these phonetically accurate, orthographic errors are the most prominent ones among spellers with no reading disabilities (Bernstein, 2009). They can therefore prove problematic for anyone relying on automatic spelling correction, regardless of learning disabilities. In this paper, we present a corpus of Icelandic homophones and a potential approach to a context sensitive spelling correction.

This paper is organized as follows: In Section 2, we discuss the task of context-sensitive spelling correction and the case of the morphologically rich Icelandic language. In Section 3, we present the compilation and contents of the Icelandic Confusion Set Corpus (ICoSIC). In Section 4, we briefly discuss our machine learning experiments with the corpus. We conclude in Section 5.

2. Context sensitive spelling correction
The idea behind the majority of spell checkers and proof-reading software commercially available is to look up an isolated word and to prompt an error message if the word doesn’t exist in the vocabulary. While this method is very useful for detecting typos and non-words, mistakes that result in real but unintended words go undetected. As noted by Rello, Ballesteros and Bigham, nearly 20% of the errors that people with dyslexia make are real-word errors and therefore it’s vital that the tools that they use can detect these spelling mistakes (Rello et al., 2015). To tackle the homophone substitution problem, another approach to spell checking is needed. Instead of looking at a word in isolation, it’s crucial to look at its context to determine which word is most likely to have been intended, given the morphological and semantic aspects of the surrounding words (Golding and Roth, 1999).

2.1. Confusion sets and the case of Icelandic
In a highly inflected language such as Icelandic, the need to disambiguate homophone word pairs is particularly apparent. Due to the morphological richness of the language, there is often very little orthographic difference between grammatical genders or cases for example which can be a great nuisance, not least for dyslexic individuals and L2 learners. As an example, the difference between the nominative and the accusative form of a masculine noun can often be found in the number of *n*’s in its suffix, i.e. *morgun* (morning, nom.) / *morgun* (morning, acc.). Another example is that the letter *y* often appears in the subjunctive past tense form of a verb, i.e. *bindi* (’bind’, subjunctive, present tense) / *byndi* (’bind’, subjunctive, past tense). As
includes a text file which contains the full list of words from separated orthographically by a single letter. Each category in the course of three months in the winter of 2019. This task was only made possible through the 2017 release of the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2019) in order to ensure that the dataset would cover as many word pairings as possible. We start by collecting sentences from the IGC which contain said word. The sentence examples are organized so that each word from the word list appears, preceded by two semicolons and followed by the appropriate sentence examples. Each line in the sentence examples contains a word and a PoS tag, separated by a tab. The confusion set categories are:

- 196 pairs containing yfí (leyti ‘extent’ / leiti ‘search’): In modern Icelandic, there is no phonetic distinction between these sounds (both of which are pronounced as [í]) and thus their distinction is purely historical. The use of y refers to a vowel mutation from another, related word, some of which are derived from Danish. Confusing words that differ only by these letters is therefore very common when writing Icelandic.
- 150 pairs containing yfí (sýn ‘vision’ / sín ‘theirs (possessive reflexive)’): The same goes for these sounds, which are both pronounced as [í]. The original rounding of y and ý started merging with the unrounded counterparts of these sounds in the 14th century and the sounds in question have remained merged since the 17th century (Gunlaugsson, 1994).
- 1203 pairs containing nn/n (forvitinn ‘curious(masc.)’ / forvitiin ‘curious (fem.)’): The alveolar nasal [n] is not elongated and therefore there is no real distinction between these sounds in pronunciation (although the preceding vowel to a double n is often elongated). The distinction between them is often grammatical and refers to whether the word has a feminine or masculine grammatical gender. However, the rules on when to write each vary and have plenty of exceptions, many of which are taught as something to remember by heart.
- 8 pairs commonly confused by Icelandic speakers: These confusion sets could prove useful in grammar correction as their difference is in their morphological information rather than their orthography. These include for example mig/mér (‘me’ (accusative) / ‘me’ (dative)) which commonly get confused when followed by experiencer-subject verbs (Jónsson and Eythórsdóttir, 2005; Ingason, 2010; Thráinsson, 2013; Nowenstein, 2017).
- 24 pairs containing hv/kv (hvad ‘what’ / kvad ‘chanted’): Hv and kv in initial position are homophones for the majority of Icelandic speakers who pronounce both as [kʰvː-]. Exceptions to this can be found in Southern Icelanders, where the initial phone is the fricative [χ] (Rögnvaldsson, 2013).
- 42 pairs containing róðr (velðri ‘weather’ (dative) / verðri ‘will become’): Included due to their potential confusability, though they are strictly speaking not homophones. These pairs are often used in tongue twisters.

3. The Icelandic Confusion Set Corpus

The focus of our research was gathering data for what has now become The Icelandic Confusion Set Corpus (hereinafter referred to as the ICoSC). It was compiled during the course of three months in the winter of 2019. This task was only made possible through the 2017 release of the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018), which consists of about 1.3 billion running words of text, tagged morphologically using IceStagger (Loftsson and Ostling, 2013). The IGC is divided into 6 text categories, including media text, official documents and the text collection of the Árni Magnússon Institute for Icelandic studies. In our project, we cross-referenced the IGC with the Database of Icelandic Morphology (Bjarnadóttir et al., 2019) in order to ensure that the dataset would cover as many word pairings as possible. We start by collecting words containing a chosen letter pair (i.e. yfí) from the DIM and then collect sentence examples and frequency information from the IGC about those pairs. The end result has been made available under a CC-BY licence on CLARIN-IS, the Icelandic repository for the European Research Infrastructure for Language Resources and Technology.

3.1. Content

The ICoSC consists of seven categories of confusion sets, selected for their linguistic properties as homophones, separated orthographically by a single letter. Each category includes a text file which contains the full list of words from that category. It also contains a text file containing all sentences from the IGC which contain said word. The sentence examples are organized so that each word from the word list appears, preceded by two semicolons and followed by the appropriate sentence examples. Each line in the sentence examples contains a word and a PoS tag, separated by a tab. The confusion set categories are:
Figure 1: Frequency table for category y/i

- 110 pairs containing rr/r (klári 'smart' (indef. fem. dative) / klári 'smart' (def. masc. nominative)): Included due to their potential confusability, as the pronunciation difference is only in the preceding vowel, similar to the nn/n-pairs.

The ICoSC also includes CSV spreadsheets which contain all the confusion sets collected for each category and their frequencies. These files are organized in the following way: for each confusion set, each candidate appears with its total frequency in the IGC. The following column shows the frequency of each possible PoS tag for the candidate in question. In the seventh and eighth column, binary values appear which refer to whether the confusion set is grammatically disjoint (the two candidates have no PoS tags in common) or grammatically identical (all PoS tags are identical for the two candidates). In the final column, the frequency of the less frequent candidate of the set is shown, which can be used to determine which sets are viable in an experiment. An example of a frequency table can be found in Figure 1.

As the n/nn examples are by far the most frequent confusion sets, the corpus also includes a word list and sentence examples for the 55 most frequent sets from that category. All files have UTF-8 encoding.

3.2. Particular uses for dyslexia in Icelandic

According to Sigurmundsdóttir and Torfadóttir (2020), learning disabilities such as dyslexia cause problems in spelling that may be even harder to attack than similar problems in reading. As people with dyslexia have a weaker phonological awareness, the conversion of sounds to orthographic symbols is often problematic. They explain that the most common symptoms of dyslexia in spelling are:

- Omission of letters.
- Difficulties distinguishing between long and short vowels. This is particularly problematic when deciding whether or not there should be a double consonant in Icelandic words, i.e. dítt (had) / dítt (ate).
- Difficulties distinguishing between voiced and unvoiced consonants, i.e. magi (stomach) / maki (romantic partner).
- Difficulties distinguishing between phonetically similar letters, i.e. dyrr (animal) / dyr (door).
- Letter switching.

As at least three of these cases can easily lead to accidental homophone mix-ups in Icelandic, a confusion set classification method is vital for the creation of a context sensitive spelling correction suitable for people with reading disabilities.

3.3. Uses for L2 learners

Another group of people that could benefit in particular from a context-sensitive proofreading software are those who are learning Icelandic as a second language. The number of immigrants living in Iceland has been steadily growing in recent years. In her 2017 pilot study, Arnórsdóttir tried to shed light on which mistakes non-native speakers are most likely to make when speaking Icelandic (Arnórsdóttir, 2017). She compared the performance of Francophones and German speakers. Her results indicate that Francophones struggle more with grammatical genders and case agreement than Germans do, indicating that language transfer might be harder from the roman languages than from other germanic languages. In any case, this indicates that L2 learners could benefit significantly from a context-sensitive spell checker.

4. Machine learning approach

After the compilation of the ICoSC, we conducted a small scale machine learning experiment on the data, using three distinct categories of confusion sets. They are:

- Grammatically disjoint word pairs (they/them): The PoS tags for each word never overlap with the other. This is very common for Icelandic. We tested 60 pairs from this category (42 taken from the n/nn category, 6 from the y/i category, 5 from the ý/í category and 7 from the various (grammatically separated) category);
- Grammatically identical word pairs (principal/principal): Both words within the pair belong to the same distributional class and differ only by semantics. Somewhat surprisingly, this turned out to be the smallest category in our research where only seven word pairs had high enough frequency to be of value (3 are from the y/i category, 2 are from the ý/í category and 2 are from the n/nn category);
- Word pairs that fall under neither aforementioned category and thus the words within the pair can differ both in their semantic and syntactic properties, (lose/loose). We tested 25 pairs from this category (8 from the n/nn category, 10 from the y/i category and 7 from the ý/í category).
The algorithm performs best on grammatically disjoint pairs, which suggest that the results could be significantly improved with a more careful consideration of the linguistic features of the context words, as they are less likely to overlap. On the other hand, the algorithm performs worst on grammatically identical pairs, where the difference between candidates is purely semantic. This could potentially be improved by looking at their semantic distance. It should be noted though that the number of grammatically identical sets is significantly lower than that of the other categories and may not be properly representative.


<table>
<thead>
<tr>
<th>Type</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disjoint</td>
<td>0.78</td>
<td>0.77</td>
<td>0.76</td>
<td>0.75</td>
</tr>
<tr>
<td>Identical</td>
<td>0.73</td>
<td>0.68</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>Overlap</td>
<td>0.79</td>
<td>0.75</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>y/i</td>
<td>0.86</td>
<td>0.76</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>y/í</td>
<td>0.79</td>
<td>0.82</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>nn/n</td>
<td>0.75</td>
<td>0.74</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>Various</td>
<td>0.75</td>
<td>0.71</td>
<td>0.66</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 1: Average scores for categories.

5. Conclusion

In recent years, Icelandic primary schools have tested children for reading disabilities within their first three months of attendance in order to ensure early intervention and that every child gets appropriate support while learning to read (Sigurmundsdóttir and Torfadóttir, 2020). The resources available for dyslexic adults are nevertheless scarce and mostly focused on reading rather than writing. No open-source spell-checking tools exist for Icelandic when this is written. The three most commonly used are Púki Writing Error Protection, Skrambi, and an Icelandic version of the Hunspell-spell checker. None of them is actually context-sensitive, although Skrambi offers a very limited confusion set lookup (Nikulásdóttir et al., 2017). However, the number of Icelandic language technology resources has finally started to grow thanks to The Icelandic language technology programme 2018-2022. It is our hope that the compilation of the ICoSC will lead to further development in context-sensitive proofreading tools, suitable for the needs of people with dyslexia and other reading disabilities.

6. Bibliographical References


