1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 1–5
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

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 minimal 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 (ICoSC). 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 proofreading 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. *morgunn* (*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 an attempt to solve this problem, a confusion set is defined consisting of word candidates that commonly get confused with one another. When a spell checker encounters these words, it tries to evaluate based on the context which candidate from the set is more likely to have been intended.

2.2. Previous work

The problem of automatically correcting real-word errors has been addressed by NLP specialists, particularly for high resource languages such as English. In their 2015 paper, Rello et al. presented a system called *Real Check*, which is based on a probabilistic language model, a statistical dependency parser and Google n-grams. They created confusion sets for Spanish using the Levenshtein Automaton dymamic algorithm in order to combat real-word errors. The results from their system is comparable to the state-of-the-art spell checkers (Rello et al., 2015). In the same year, Rokaya used a combination of the confusion set method and statistical methods to disambiguate semantic errors in Arabic (Rokaya, 2015) and Samani M.H., Rahimi Z. and Rahimi S. addressed real-word spelling mistakes in Persian using n-gram based context retrieval for confusion sets (Samani et al., 2015). Both experiments resulted in around 85-90% precision rate. In the case of Icelandic, Ingason et al. conducted a small-scale experiment in 2009 using feature extraction from the context of confusion set candidates. These features were then fed to the Naive Bayes and Winnow algorithms with promising results. We hope to expand this research in our experiments, using a much larger database than previously available.

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 Östling, 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. *y/i*) 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:

- 196 pairs containing y/i (*leyti 'extent' / leiti 'search'*): In modern Icelandic, there is no phonetic distinction between these sounds (both of which are pronounced as [ɪ]) 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.

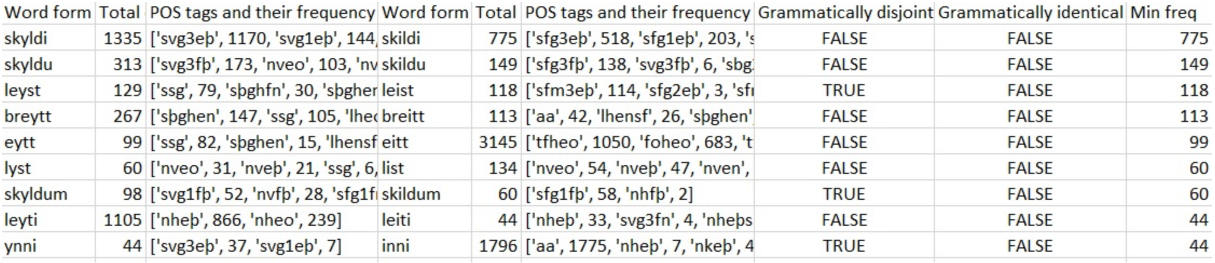
- 150 pairs containing ý/í (*sýn 'vision' / sín 'theirs (possessive reflexive)'*): The same goes for these sounds, which are both pronounced as [i]. The original rounding of y and ý started merging with the unrounded counterparts of these sounds in the 14th century and the sounds in question have remained merged since the 17th century (Gunnlaugsson, 1994).

- 1203 pairs containing nn/n (*forvitinn 'curious(masc.)' / forvitin '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. It is therefore common for both native and nonnative speakers to make spelling and/or grammar mistakes in these type of words.

- 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 Eythórsson, 2005; Ingason, 2010; Thráinsson, 2013; Nowenstein, 2017).

- 24 pairs containing hv/kv (*hvað 'what' / kvað 'chanted'*): Hv and kv in initial position are homophones for the majority of Icelandic speakers who pronounce both as [k^hv-]. Exceptions to this can be found in Southern Icelanders, where the initial phone is the fricative [x] (Rögnvaldsson, 2013).

- 42 pairs containing rð/ðr (*veðri 'weather' (dative) / verði 'will become'*): Included due to their potential confusability, though they are strictly speaking not homophones. These pairs are often used in tongue twisters.



Word form	Total	POS tags and their frequency	Word form	Total	POS tags and their frequency	Grammatically disjoint	Grammatically identical	Min freq
skyldi	1335	['svg3eþ', 1170, 'svg1eþ', 144,	skildi	775	['sfg3eþ', 518, 'sfg1eþ', 203, 's	FALSE	FALSE	775
skyldu	313	['svg3fþ', 173, 'nveo', 103, 'nv	skildu	149	['sfg3fþ', 138, 'svg3fþ', 6, 'sbg	FALSE	FALSE	149
leyst	129	['ssg', 79, 'sþghfn', 30, 'sþgher	leist	118	['sfm3eþ', 114, 'sfg2eþ', 3, 'sfı	TRUE	FALSE	118
breytt	267	['sþghen', 147, 'ssg', 105, 'lhe	breitt	113	['aa', 42, 'lhensf', 26, 'sþghen'	FALSE	FALSE	113
eytt	99	['ssg', 82, 'sþghen', 15, 'lhensf	eitt	3145	['tfheo', 1050, 'foheo', 683, 't	FALSE	FALSE	99
lyst	60	['nveo', 31, 'nveþ', 21, 'ssg', 6,	list	134	['nveo', 54, 'nveþ', 47, 'nven',	FALSE	FALSE	60
skyldum	98	['svg1fþ', 52, 'nvfþ', 28, 'sfg1f	skildum	60	['sfg1fþ', 58, 'nhfþ', 2]	TRUE	FALSE	60
leyti	1105	['nheþ', 866, 'nheo', 239]	leiti	44	['nheþ', 33, 'svg3fn', 4, 'nheþs	FALSE	FALSE	44
ynni	44	['svg3eþ', 37, 'svg1eþ', 7]	inni	1796	['aa', 1775, 'nheþ', 7, 'nkeþ', 4	TRUE	FALSE	44

Figure 1: Frequency table for category y/i

- 110 pairs containing rr/r (*klárri '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 eight 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. *áttu (had)* / *átu* (ate).
- Difficulties distinguishing between voiced and unvoiced consonants, i.e. *magi (stomach)* / *maki (romantic partner)*.
- Difficulties distinguishing between phonetically similar letters, i.e. *dýr (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 Francophone 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 *(principle/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.

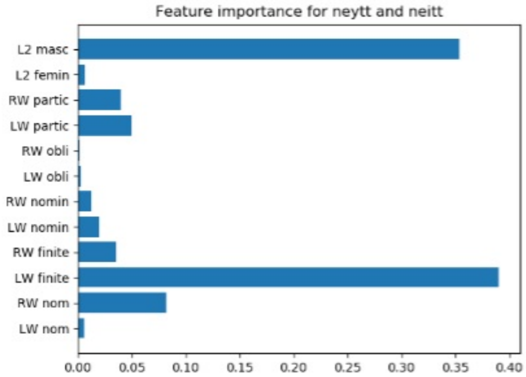


Figure 2: Feature importance for neytt 'consumed' / neitt 'anything'

In our experiment, first published in Friðriksdóttir and Ingason 2020, we used the decision tree algorithm from Scikit learn (Pedregosa et al., 2011) to create a binary classifier. We extracted linguistic features from the context of the confusion set candidates, taking into consideration the two closest words to the left of the candidate as well as the single closest word to the right of the candidate. As Icelandic grammar is quite regular, the presence of a finite verb for example can give a lot of important grammatical information of the neighbor word. We chose this narrow context for its simplicity, but adding the second word to the left is intended to capture the subject of the phrase (i.e. *"**he** is happy"* or *"**the girl** is running"*). The features were handpicked by the authors for their assumed generalizability and have binary values (true/false). The following were considered for both left and right context words: is nominal (words with grammatical case, such as nouns and pronouns); is finite (a verb that inflects for person agreement); is nominative; is oblique (has some grammatical case other than nominative); is a particle. For the word second to the left of the target word we consider if it is feminine or masculine. Example of the feature importance for a specific confusion set can be seen in Figure 2. The results were obtained using 10-fold cross validation on all the sentence examples in the data containing the two candidates. While our experiment should be considered as proof of concept rather than a finalized result, the average precision obtained for all categories ranged from 73-86% (see Table 1 which includes average for all word pairs taken from the two types of categories), indicating that results could be perfected with further research.

Type	Accuracy	Precision	Recall	F-score
Disjoint	0.78	0.77	0.76	0.75
Identical	0.73	0.68	0.66	0.64
Overlap	0.79	0.75	0.68	0.68
y/i	0.86	0.76	0.74	0.73
ý/í	0.79	0.82	0.79	0.78
nn/n	0.75	0.74	0.73	0.70
Various	0.75	0.71	0.66	0.66

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

Arnórsdóttir, A. L. (2017). *Je parle très bien l'islandais, surtout à l'écrit: recherche sur les transferts du français vers l'islandais chez les apprenants francophones*. Unpublished BA-thesis, University of Iceland.

Bernstein, S. E. (2009). Phonology, decoding, and lexical compensation in vowel spelling errors made by children with dyslexia. *Reading and Writing*, 22(3):307–331.

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, Turku, Finland.

Friðriksdóttir, S. R. and Ingason, A. K. (2020). Disambiguating confusion sets in a language with rich morphology. In *The 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, number 1, pages 446–451.

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

Gunnlaugsson, G. M. (1994). *Um afkringingu á/y, ý, ey/í íslensku*. Málvísindastofnun Háskóla Íslands.

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

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

Loftsson, H. and Östling, R. (2013). Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 105–119, Oslo, Norway, May. Linköping University Electronic Press, Sweden.

Mody, M. and Silliman, E. R. (2008). *Brain, behavior, and learning in language and reading disorders*. Guilford Press.

Nikulásdóttir, A. B., Guðnason, J., and Steingrímsson, S. (2017). *Language Technology for Icelandic. Project Plan*. Icelandic Ministry of Science, Culture and Education.

Nowenstein, I. (2017). Determining the nature of intraspeaker subject case variation. In Caroline Heycock Hjalmar P. Petersen Thráinsson, Höskuldur et al., editors, *Syntactic Variation in Insular Scandinavian*, pages 91–112. John Benjamins.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rello, L., Ballesteros, M., and Bigham, J. P. (2015). A spellchecker for dyslexia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 39–47.

Rögnvaldsson, E. (2013). Hljóðkerfi og orðhlutakerfi íslensku. *Reykjavík: Eiríkur Rögnvaldsson. Link: https://notendur. hi. is/eirikur/hoi. pdf.*

Rokaya, M. (2015). Arabic semantic spell checking based on power links. *International Information Institute (Tokyo). Information*, 18(11):4749–4770, 11.

Samani, M. H., Rahimi, Z., and Rahimi, S. (2015). A content-based method for persian real-word spell checking. In *2015 7th Conference on Information and Knowledge Technology (IKT)*, pages 1–5, May.

Sigurmundsdóttir, H. and Torfadóttir, S. (2020). Lesvefurinn um læsi og lestrarerfiðleika. Last accessed: February 7th 2020.

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.

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