Post-Correction of Icelandic OCR Text

Jón Friðrik Daðason

**Faculty of Industrial Engineering,
Mechanical Engineering and
Computer Science
University of Iceland**

2012

Post-Correction of Icelandic OCR Text

Jón Friðrik Daðason

60 ECTS thesis submitted in partial fulfillment of a
Magister Scientiarum degree in Computer Science

Advisors
Sven Þ. Sigurðsson
Kristín Bjarnadóttir

Faculty Representative
Hrafn Loftsson

Faculty of Industrial Engineering, Mechanical Engineering and
Computer Science
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, June 2012

Post-Correction of Icelandic OCR Text
60 ECTS thesis submitted in partial fulfillment of a *Magister Scientiarum* degree in Computer Science

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Hjarðarhagi 2-6
107, Reykjavik
Iceland

Telephone: 525 4000

Bibliographic information:
Jón Friðrik Daðason, 2012, *Post-Correction of Icelandic OCR Text*, Master's thesis, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland.

Printing: Samskipti ehf.
Reykjavik, Iceland, June 2012

Abstract

The topic of this thesis is the post-correction of Icelandic OCR (optical character recognized) text. Two methods for spelling correction of OCR errors in Icelandic text are proposed and evaluated on misrecognized words in a digitization project which is ongoing in Alþingi (the Icelandic parliament). The first method is based on a noisy channel model. This method is applied to nonword errors, i.e., words which have been misrecognized during the OCR process and transformed into another word which is not in the Icelandic vocabulary. This method achieves a correction accuracy of 92.9% when applied to a test set of nonword errors from a large collection of digitized parliamentary speeches from the Alþingi digitization project (a total of 47 million running words from the years 1959-1988). The second method uses Winnow classifiers, and is applied to real-word errors, i.e., words which have been misrecognized during the OCR process and transformed into another word which also exists in the Icelandic vocabulary. A Winnow classifier is able to correct real-word errors by detecting words which do not fit in the context in which they appear and suggesting other similar words which are more likely to be correct. When applied to a test set of real-word errors from the same set of digitized texts as above, this method achieves a correction ratio of 78.4%. When both methods are applied to all errors in the digitized parliamentary speeches, an overall correction accuracy of 92.0% is achieved.

Útdráttur

Efni þessa verkefnis er leiðrétting á ljóslesnum (e. optical character recognized, OCR) íslenskum texta. Tvær aðferðir til að leiðrétta ljóslestrarvillur í íslenskum texta eru þróaðar og síðan metnar á villum í ljóslestrarverkefni sem Alþingi stendur að. Fyrri aðferðin byggir á líkani fyrir leiðréttingu á stafsetningarvillum sem orsakast af truflunum í samskiptarásum (e. noisy channel spelling correction). Hún er notuð til að leiðrétta ósamhengisháðar villur, þ.e. villur þar sem ljóslestur á orði misheppnast þannig að orð breytist í annað orð sem ekki er til í íslensku. Með þessari aðferð reynist unnt að leiðrétta 92,9% af slíkum villum þegar aðferðinni er beitt á safn ljóslesinna þingræðna úr Alþingisverkefninu (samtals 47 milljón lesmálsorð frá árunum 1959-1988). Seinni aðferðin notar vélrænan Winnow flokkara og er beitt á samhengisháðar villur, þ.e. orð þar sem ljóslestur hefur mistekist þannig að orð breytist í annað orð sem þó er til í íslensku. Winnow flokkari getur leiðrétt slíkar ljóslestrarvillur með því að finna orð sem falla ekki að því samhengi sem þau koma fyrir í og stinga upp á öðrum orðum í staðinn sem eru líklegri til að vera rétt. Með þeirri aðferð tekst að lagfæra 78,4% af öllum samhengisháðum villum í þessu sama textasafni. Þegar báðum aðferðunum er beitt á þetta safn reynist unnt að lagfæra 92,0% af öllum villum.

Table of Contents

List of Figures

List of Tables

Acknowledgements

In the summer of 2010, I worked on a two-month project at the Árni Magnússon Institute for Icelandic Studies (ÁMI) under the guidance of Kristín Bjarnadóttir, Sigrún Helgadóttir and Ásta Svavarsdóttir. This project involved the post-correction of digitized 19th century periodicals in Icelandic and was my introduction to the field of natural language processing. A year later, I received a grant from the Icelandic Student Innovation Fund (ISIF) to implement and evaluate a selection of methods for context-sensitive spelling correction for Icelandic. I worked on this project during the summer of 2011, under the supervision of Sven Þ. Sigurðsson and Kristín Bjarnadóttir, in facilities kindly provided by the ÁMI. This project received a recognition of excellence from the ISIF board. This thesis is in many respects a continuation of these two earlier projects.

Special thanks go to my advisors, Sven Þ. Sigurðsson and Kristín Bjarnadóttir, for their invaluable guidance and support over the course of this project. I would also like to thank Sigrún Helgadóttir for her helpful advice and suggestions, and Hrafn Loftsson for his contributions as faculty representative.

A part of this thesis, a tool to identify and analyze Icelandic compound words, was completed as the final project of the graduate course Computers and Language, taught by Eiríkur Rögnvaldsson at the University of Iceland, to whom I would like to express my thanks. The creation of this tool was made possible thanks to unpublished data on Icelandic compounds kindly provided by Kristín Bjarnadóttir, for which I am very grateful.

I would also like to express my gratitude to Sigurður Jónsson for generously providing me with an extensive collection of digitized texts from an ongoing digitization project at Alþingi, the national parliament of Iceland. Furthermore, I am also very grateful to the employees of Alþingi who have been working on digitizing and proofreading these documents over the last ten years. The data resulting from their hard work was invaluable to this research.

I would also like to express my sincere thanks to the Árni Magnússon Institute for Icelandic Studies and its employees for all the help they have given me over the last few years, without which this thesis would not exist. My thanks also go to Kristján Rúnarsson who has worked with me on previous projects, including the OCR post-correction project in the summer of 2010.

Finally, I would like to thank my family for their continued support and encouragement.

1 Introduction

Enormous quantities of printed media such as books, newspapers and periodicals are not available in an electronic text format. Their manuscripts may have been written before the advent of word processing and never have been digitized, or have been lost over time or stored on antiquated hardware or in obsolete formats.

Many institutions, corporations and volunteers are and have been engaged in large-scale digitization projects. Probably the most ambitious of these is one led by Google (as part of their Google Books service), which aims to digitize all books which have ever been published by the end of the decade (Orwant 2010). As of March 2012, they have digitized over 20 million books (Howard 2012) out of what they estimate to be around 130 million in total (Taycher 2010). Other projects include the volunteer-based Project Gutenberg and Microsoft's Live Search Books service.

The main goal of these digitization projects is to make non-digital printed media widely available, distributable and searchable online. These digitized texts are a rich resource both for historical and linguistic research, as well as for the public at large who may have an interest in their contents.

However, even under optimal conditions, it is inevitable that some errors will be introduced during the optical character recognition (OCR) process. The digitized text may require proofreading if its correctness is highly valued.

1.1 Scope of this Thesis

The goal of this thesis is to investigate the challenges presented by the correction of OCR errors in large-scale digitization of Icelandic text. Specifically, this thesis will focus on the Alþingi digitization project.

Alþingi, the national parliament of Iceland, has been digitizing printed copies of its parliamentary records since March 2002. The records in question are composed of approximately 185,000 pages of speeches and 190,000 pages of parliamentary documents from the years 1845-1992. The printed records are scanned and then processed by OCR software. The parliamentary document portion is then immediately made available online, while the speech portion first undergoes proofreading. Parliamentary speeches in their proofread and unproofread forms from the years 1959-1988 (a total of 47 million running words in each case) have been made available for the purposes of this research.

The research question is threefold: What are the main challenges presented by Icelandic with regard to spelling correction, how well do established methods for spelling

correction apply to the correction of Icelandic text, and how well do they apply to the correction of OCR errors?

1.2 Structure of this Thesis

Chapter 2 contains a discussion of the nature of OCR errors and an overview of the main challenges involved in correcting Icelandic text. The differences between OCR errors and human spelling errors are examined and a brief summary of several different methods for correcting spelling errors is given. Additionally, several Icelandic language resources which are relevant to this project are introduced.

A more in-depth analysis of the various methods for spelling correction as they relate to the task at hand is given in chapter 3. Methods which are found to be the most likely to yield good results when applied to the digitized Alþingi texts are selected for evaluation. Furthermore, some steps to overcome the challenges posed by correcting Icelandic text are introduced. This chapter is split into three parts: The first introduces a method to deal with issues related to the propensity of compound words in Icelandic, the second considers various methods for correcting spelling errors which do not make use of the context in which they appear, and the third discusses correction methods which do make use of contextual information.

These methods are evaluated on the digitized Alþingi texts in chapter 4.

Chapter 5 summarizes the main results of this thesis and discusses future work.

An overview of the most common OCR errors introduced to the digitized parliamentary speeches is given in appendices A and B.

1.3 User Interface

As stated in the research question, the primary goal of this thesis is to identify and evaluate methods for spelling correction which are viable for the task of correcting digitized Icelandic text. Delivering a real-world application, i.e., a fully-featured spellchecker with a user interface is not one of its goals. However, there are some benefits to be had from combining all the components of the spellchecker under a user interface which are worth mentioning.

A user interface makes it easy for the spellchecker to give feedback to the user and vice-versa. For example, it becomes a simple task for a user to ignore well-spelled words which are mistakenly identified as errors. Without intervention from the user, the spellchecker might attempt to “correct” such false positives. It also allows for multiple correction candidates to be suggested by the spellchecker, ranked by their probability of being correct. This means that the accuracy of the spellchecker is no longer a question of how often the most likely word is correct, but rather how often it is among the five or so most likely candidates. Additionally, it becomes easier to dynamically update the spellchecker’s

error model to reflect recently corrected errors. In this manner, the spellchecker might adapt its error model to fit a particular user or a document.

Even though a user interface will not be developed as part of this thesis, the methods will still be evaluated with real-world applicability in mind. As such, it is always assumed that there is a user operating the spellchecker, and that he will be presented with false positives.

1.4 Implementation

The source code for the methods evaluated in this work is written in the Python programming language, and will be made publically available shortly following the publication of this work. Two third-party libraries and toolkits were used in this project. The IceNLP toolkit (Loftsson and Rögnvaldsson 2007) is used for tokenization, lemmatization and part-of-speech tagging. The Diff, Match and Patch Library (Fraser 2012) is used in order to attain a minimal difference between two strings.

2 Foundations

This chapter gives an overview of the errors which can occur during the digitization of text. The challenges involved in digitizing Icelandic text are considered in particular. Certain qualities of Icelandic which may complicate the post-correction of digitized text are identified and discussed. Finally, some language resources which are relevant to the task of correcting Icelandic text are introduced.

2.1 Optical Character Recognition

Optical character recognition (OCR) is an automated process by which text is extracted from an image. The purpose of this work is to improve the output of this process, as opposed to improving the process itself.

Many OCR software vendors claim an accuracy rating of over 99% under optimal conditions. This assumes that the document undergoing OCR is relatively recent and that the image is clean and of high quality. The National Library of Australia tested the character accuracy of a sample of 45 pages of digitized newspapers from the years 1803-1954 and found that for each page, it was between 71-98% (Holley 2009). The National Library of the Netherlands conducted a survey among companies involved in large-scale newspaper digitization projects. It reports that a digitization project of 350,000 pages of 20th century newspapers had a character accuracy of 68% (Klijn 2008). This shows that character accuracy can fall dramatically under non-optimal settings.

The parliamentary speeches from the Alþingi digitization project which were made available for this thesis are quite recent, having been printed in the latter half of the 20th century. A comparison of the proofread speeches and their unproofread counterparts shows that the OCR software achieved an average character accuracy of 99.2%.

2.1.1 OCR and Icelandic

Developers of OCR software may lack the incentive to support languages with few speakers if their alphabet or script differs from those of more commonly spoken languages. Even if they are supported, substantial effort may be required on part of the developer to ensure that the character accuracy for such languages does not fall far behind that of the most commonly spoken languages. In the face of such costs, developers may be willing to accept worse performance for languages with few speakers.

The Icelandic alphabet consists of 32 letters, including several with diacritics (*á*, *é*, *í*, *ó*, *ú*, *ý* and *ö*) as well as the letters *ð*, *þ* and *æ*. In total, it contains 10 non-English alphabetic characters, which account for roughly 15% of all Icelandic alphabet characters occurring in the proofread parliamentary speeches. Despite their relatively low frequency, OCR errors involving at least one of these characters were responsible for over 56% of all

words[1] which were misrecognized by the OCR software. This means that the majority of word errors which were corrected by proofreaders were caused by an OCR error where a non-English alphabet character was misrecognized as some other character (e.g., where *þ* was mistaken for *p*), or some character was misrecognized for a non-English alphabetical character (e.g., where *i* was mistaken for *í*).

The fact that these relatively infrequent characters are responsible for the majority of word errors implies that Icelandic is at a disadvantage compared to other more commonly spoken languages when it comes to OCR accuracy.

2.1.2 Comparison of Human Errors and OCR Errors

People can make a wide variety of different errors when writing text. These errors can be typographical in nature (e.g., if *teh* is written instead of *the*), grammatical (e.g., writing *to who it may concern* instead of *to whom it may concern*) or they can be spelling errors (e.g., confusing *lose* with *loose*). In comparison, most OCR errors are due to characters in an input document being mistaken for others whose shape is similar.

This can give the appearance that OCR errors are more random in nature than human errors. Indeed, it can be difficult to tell beforehand what kind of errors will occur during the OCR process. The use of different OCR software or even a subtle difference in the font a document uses might make the difference between *þ* being the most or the least commonly misrecognized character.

However, OCR errors are not random in the sense that it is extremely unlikely for the letter *o* to be misrecognized as *x*. The two characters are simply too dissimilar in appearance. There is always a reason for the mistakes the OCR software makes, and due to this reason, it will always repeat it under the exact same settings. To expand on this point, one could imagine a page representative of a digitized book. If the page contains a number of errors where *m* has been misrecognized as *rn*, it is likely that this error is common on other pages as well. Similarly, if the letter *a* appears many times on that page yet is never misrecognized, then most occurrences of that letter on other pages are likely to be correct as well.

2.1.3 Categories of OCR Errors

Character Errors

During the OCR process, characters in the input document may be replaced with other characters, often ones which are similar in appearance. For example, the character *d* might be replaced with *cl* and *m* might be replaced with *rn*. Characters may also be missed entirely, for example because of faded text. It is also possible for characters to be inserted into the OCR generated text without the presence of a corresponding character in the input document. This may occur when images or smudges are mistaken for text or if characters from one side of the page can be seen on the opposite side (an effect called

[1] Here, *words* refers to all words occurring in the text, including duplicates, (also known as *running words* or *tokens*), as opposed to *word forms* (or *types*) which refers only to distinct running words. In this work, a *word* may refer to either a running word or a word form when the meaning is clear from the context.

bleed-through). All of these errors are known as character errors. More precisely, a character error is every single edit operation (a deletion, an insertion or a replacement) which must be applied to the OCR generated text in order to make it identical to the input document.

Word Errors

A word error is a word in the input document which is not present and correct in the OCR generated text. These words are said to have been misrecognized. Word errors can be caused by a character error in the word itself or by mistakenly joining two or more words. The latter possibility is known as a word boundary error. Such errors come in two varieties, incorrect splits where a whitespace character has been inserted into a word, and run-ons where whitespace characters separating two words have been removed. Word errors may also occur when a word is missed entirely, usually caused by zoning errors.

Zoning Errors

A page may be composed of one or more zones, each containing a block of contiguous text. A page with two columns of text therefore contains two zones, and a page split into two columns and two rows contains four zones. It is up to the OCR software and/or its users to correctly identify the zones and the order in which they are meant to be read.

A zoning error refers to one of three things:

- Text which has not been identified by the OCR software as being inside any zone, causing it to be missing from the output.
- Non-contiguous text which has been determined by the OCR software as belonging to the same zone, resulting in text appearing out of order in the output. An example of this is when two side-by-side columns of text are placed within the same zone. This would result in each line of the latter column being appended to the corresponding line of the former column in the OCR generated text.
- The incorrect ordering of zones, resulting in text appearing out of order in the output. An example of this is when a page is split into two columns and two rows. A zoning error would occur if text were output in a row-by-row order when it was meant to be read in a column-by-column order.

2.2 Error Detection and Correction

2.2.1 Nonword Errors

A word is said to be a nonword in a particular language if it is not recognized as a well-spelled word in its vocabulary. If a misspelled word is a nonword it is known as a nonword error. An example of such errors is when *taka* ‘take’ is misrecognized as *faka*, a nonword in Icelandic, or *þeir* ‘they’ as the nonword *peir*.

One method of identifying nonword errors is by searching for sequences of n characters (character n-grams) which are highly unlikely or not known to appear within well-spelled

words. Should the character 3-grams *yii*, *tfs* or *vlð* appear within a word, it is almost certainly not a well-spelled Icelandic word and can be safely flagged as a nonword error. However, as the 3-grams *har*, *ari* and *ris* are all common in Icelandic, the nonword *haris* would not appear to be suspicious at all and would not be flagged.

Another method of detecting nonword errors is by the use of a lexicon (often referred to as a dictionary and sometimes as a word list in the context of spellchecking). In this approach each word form in a document is compared to a list of known well-spelled word forms. Word forms not present in the list are flagged as nonword errors. The accuracy of this method depends on the lexicon's coverage of the vocabulary and the amount of undesirable words it contains. An undesirable word could be a nonword or an obsolete or very rare word, an occurrence of which is far more likely to be a misspelling of another word than a legitimate usage. One such example is the word *vili* 'will', an obsolete variation of the more modern word *vilji*. It also exists as a proper noun, being the name of a minor figure in Norse mythology. Appearing in OCR text, it is almost certainly an error, most likely a misrecognition of the word *vill* 'wants'. Even though in some sense *vili* is a part of the vocabulary of Icelandic, including it in the lexicon would probably do more harm than good.

The construction of the lexicon is a balancing act where the benefits of increased coverage must be weighed against the disadvantage of adding undesirable words. A lexicon constructed from a traditional dictionary for human consumption can have fair coverage and a very low number of unwanted words. On the other hand, a lexicon constructed from a very large collection of non-proofread text is likely to have very high coverage but also a relatively high number of undesirable words. The first lexicon might cause many well-spelled words to be flagged as nonwords and the second might cause many nonwords to be accepted as being well-spelled.

Once a nonword error has been detected, the spell checker may suggest alternatives to the user. This can be accomplished with the aid of a lexicon. Possible candidates for correction would be words within a certain edit distance of the nonword. An edit distance is a metric for calculating the difference between two words. One example of such a metric is the Levenshtein distance (Levenshtein 1966), which considers every deletion or insertion of a single character or replacement of one character with another to be a single edit. More information may be used to improve the accuracy of the suggestions, including the word frequency of each correction candidate or the probability of the error occurring. For example, it is much more likely that OCR software will mistake *i* for *l* than *x* for *o*. This is discussed in greater detail in section 3.2.

2.2.2 Real-word Errors

If a misspelled word also happens to be a valid, well-spelled word, it is said to be a real-word error. An example of such errors is when *deila* 'dispute' is misrecognized as *della* 'nonsense' or *línur* 'lines' as *linur* 'soft'. Because real-word errors only involve well-spelled words, they are impossible to detect by examining one word at a time. The error can only be revealed by considering the context in which the word appears. For example, *gefa* 'give' by itself gives no indication that it is a misspelling, but it is clearly an error if it

appears in the sentence “Hvernig gefa fuglar flogið?” (“How give birds fly?”). In fact, it is quite clear from the context that the sentence should instead be “Hvernig geta fuglar flogið?” (“How can birds fly?”). Since the correction of real-word errors depends on their context, the act of correcting them is known as context-sensitive spelling correction.

Correction as a disambiguation problem

Context-sensitive spelling correction may be viewed as a disambiguation problem. In this approach, words which are likely to appear as misspellings of one another are grouped together in so-called confusion sets. If a word which belongs to some confusion set appears in a sentence, it is considered to be ambiguous whether it is actually correct or if it is a misspelling of another word from the same set. Disambiguation means determining which word in the confusion set is the likeliest to be correct, given the surrounding context. This can be achieved by training a spell checker on a large corpus of text in order to learn in which context each word in the confusion set is likely to appear. Employing this method with naive Bayesian (Golding 1995) and Winnow classifiers (Golding and Roth 1999) has given good results. These methods are discussed in greater detail in section 3.3.

Rule-based correction

Another way of correcting real-word errors is by constructing rules. A rule might state that every occurrence of *skýrar linur* ‘clear soft’ should be replaced by *skýrar línur* ‘clear lines’. Creating the rules by hand on a large scale is obviously an infeasible task, but methods for automatic rule acquisition have proven to be quite effective (Mangu and Brill 1997).

Correction with statistical language models

It is also possible to detect real-word errors by examining all sequences of n words (word n-grams) within a sentence (Jurafsky and Martin 2009). For this to be possible, a statistical language model is required. The model will estimate each n-gram’s probability of appearance, based on how often it occurs within a large corpus. An n-gram which is not present in the corpus or has a very low probability of appearing, indicates that it may contain an error. Likely candidates for correction are n-grams which are a short edit distance away with a high probability of appearance.

2.3 Challenges

Certain features of Icelandic complicate research and development in the field of language technology. These features are both of a geopolitical and grammatical nature.

2.3.1 Geopolitical

Iceland is a country with a population of approximately 320,000. When it comes to language technology, small language communities are at a disadvantage for several reasons. The abundance of resources (see section 2.3.4) which are available for a given language in part depends on the number of its speakers. The fewer the speakers, the scarcer the resources tend to be. Furthermore, the cost of creating language resources

for a particular language is not affected by the number of its speakers. Therefore, the cost per speaker becomes higher the fewer they are.

2.3.2 Rich Morphology

Inflection

In an inflectional language, grammatical categories such case, number, tense, gender, etc. are exhibited by different word forms, i.e., inflectional forms. For example, the plural of *maður* 'man' is *menn* 'men' and the past tense of *halda* 'hold' is *hélt* 'held'. Not all languages are inflectional (e.g., Mandarin Chinese), and, those that are, vary in the degree to which inflection is present. For example, English is a weakly inflectional language while Icelandic is highly inflectional.

To name an example, inflectional categories for English nouns are case (nominative and genitive) and number (singular and plural), resulting in a maximum of 4 inflectional forms, excluding variants. In comparison, the inflectional categories for Icelandic nouns are case (nominative, accusative, dative and genitive), number (singular and plural), and definiteness, i.e., the presence or absence of the suffixed definite article. The maximum number of inflectional forms for Icelandic nouns is therefore 16, excluding variants (Bjarnadóttir 2012a).

Table 2.1 The inflectional paradigm for the word maður *'man'*

	Singular			**Plural**	
	Indefinite	**Definite**		**Indefinite**	**Definite**
Nom.	maður	maðurinn	**Nom.**	menn	mennirnir
Acc.	mann	manninn	**Acc.**	menn	mennina
Dat.	manni	manninum	**Dat.**	mönnum	mönnunum
Gen.	manns	mannsins	**Gen.**	manna	mannanna

Table 2.1 shows all 16 inflectional forms of the noun *maður* 'man'. The word form *maður* is known as the lemma (also called headword or dictionary form) of these inflectional forms. Two of the inflectional forms are identical, meaning that there are a total of 15 word forms among them. Adding the word to a spellchecker's lexicon means that all 15 word forms must be added. On the other hand, the English equivalent, *man*, only has 4 inflectional forms: *man*, *men*, *man's* and *men's*. All other things being equal, a highly inflectional language probably requires a substantially larger lexicon to maintain the same degree of coverage of its vocabulary as a weakly inflectional language.

Constructing a lexicon is more problematic in a highly inflectional language. Dictionaries are less helpful as source material as they only tend to list words by their lemmas and provide limited information on the inflection. Large collections of text might be more suitable as source material, but the more inflective a language is, the larger the collection of text will have to be. Also, large collections of text inevitably contain some spelling errors and other unwanted word forms.

A related issue is one of a statistical nature. One way to explain it is to imagine a faithful translation of an English text to Icelandic (or vice versa). Assuming that the translation does not stray too far from the original, it would not be farfetched to assume that the English word *man* and the Icelandic word *maður* should occur with roughly the same frequency in both versions.

Table 2.2 Comparison of the frequency of maður *and* man *in Lovecraft's Call of Cthulhu*

English		**Icelandic**							
33	men	10	menn	5	mönnum	3	mannsins	1	mannanna
21	man	9	maður	4	manni	3	mennirnir	1	manninum
3	man's	8	manna	3	mann	2	manninn	1	mennina
1	men's	5	maðurinn	3	manns	2	mönnunum		

Table 2.2 shows the frequency of each inflectional form of *man* in the short story *The Call of Cthulhu* (Lovecraft 1926) and of *maður* in its Icelandic translation (Lovecraft 2011). The words occur with roughly the same frequency in both version (58 times in English and 60 times in Icelandic), but the effects that inflection has are clearly visible. Every occurrence in English takes one of four word forms, while every occurrence in Icelandic takes one of 15. It seems from these numbers that a word frequency analysis (i.e., word form frequency analysis) gives more valuable results when performed on the English text than its Icelandic translation.

On average, there are over 25 word forms among the inflectional forms of verbs in the Database of Modern Icelandic Inflection (see section 2.3.4). How often must an Icelandic verb occur in a given text before all of its inflectional forms have appeared at least once? How often before it becomes possible to accurately estimate each word form's probability of appearance? On the other hand, weak English verbs have a total of four inflectional forms, while strong verbs have five. It seems likely that a greater amount of text is required to create a useful statistical language model for Icelandic than for English.

Compounds

Icelandic is a compounding language, which means that two or more words can be combined to form a new word, a compound or compound word. Examples of these are *hljóðbylgja* 'sound wave' (*hljóð* 'sound' + *bylgja* 'wave') and *smásaga* 'short story' (*smá* 'short' + *saga* 'story'). While there are certain constraints as to how compounds can be formed, there is theoretically no limit to how many words can be combined to form a single compound. This makes it possible to create long compound words such as *kjarnorkuendurvinnslustöð* 'nuclear recycling plant' and *Alþjóðaheilbrigðismálastofnunin* 'World Health Organization' (Bjarnadóttir 2002).

When it comes to spelling correction, there are two possible strategies that can be taken to handle compounds. The first is simply to add as many of them as possible to the lexicon. This option may be infeasible for languages in which they are common and will possibly result in the lexicon multiplying in size. The other option is to attempt to identify whether an unknown word is a compound or not by considering which words it might be

composed of, a process known as decompounding. The downside of decompounding is that it may end up accepting compounds which are either not well-formed or meaningless. Decompounding is discussed further in section 3.1.

2.3.3 Word Order

Icelandic has fewer constraints on word order than some other languages, as can be seen in the following example (Rögnvaldsson 1990:60):

- María$_{<nom.subj.>}$ kyssti Svein$_{<acc.obj.>}$ (Mary$_{<sub.>}$ kissed Sven$_{<subj.>}$)
- Svein$_{<acc.obj.>}$ kyssti María$_{<nom.subj.>}$ (Sven$_{<obj.>}$ kissed Mary$_{<subj.>}$)
- cf. Sveinn$_{<nom.subj.>}$ kyssti Maríu$_{<acc.obj.>}$ (Sven$_{<sub.>}$ kissed Mary$_{<obj.>}$)

One effect of this increased flexibility is that word n-grams which might otherwise have no chance of appearing become possible. Another is that the frequency of a certain sentence will be divided up between all the possible ways in which its words can be ordered. These issues probably mean that the fewer constraints there are on word order in a particular language, the greater the amount of text is needed to create a good statistical language model for it.

2.3.4 Language Resources

The Open Language Archive Community defines a language resource as “any physical or digital item that is a product of language documentation, description, or development or is a tool that specifically supports the creation and use of such products.” (Simons and Bird 2008)

One example of a language resource is a corpus, a large digital collection of text or speech which is generally chosen to meet certain criteria. A text corpus might be constructed with the aim of representing contemporary usage of some language, in which case its texts will ideally be relatively recent and from a wide variety of genres (to reduce bias). The corpus may be annotated (or tagged), for example by including grammatical information for each word. Such a corpus would be valuable both for general language research as well as for the creation and/or evaluation of grammatical analysis tools.

Languages in which resources are scarce are known as less-resourced languages (or under-resourced languages). There is no clear metric which can be used to measure the abundance and quality of a language’s resources, making this a subjective classification. In META-NET’s language report for Icelandic (Rögnvaldsson et al. 2012), the language was categorized as having weak or no support for language resources (along with Irish, Latvian, Lithuanian and Maltese). By most definitions, this would be enough to justify Icelandic being classified as a less-resourced language. However, the report also states that “Icelandic stands reasonably well with respect to the most basic language technology tools and resources, such as text analysis and text corpora.”

Despite its status as a less-resourced language, there are many Icelandic language resources which are relevant to spelling correction in the context of the work described in this thesis. The following is a list of resources which will be used in this work:

IceNLP

IceNLP (Loftsson and Rögnvaldsson 2007) is a natural language processing toolkit for Icelandic. It includes a part-of-speech (POS) tagger capable of grammatically analyzing Icelandic text and a lemmatizer which attempts to determine a word's lemma. The POS-tagger has an estimated accuracy of 92.51% (Loftsson et al. 2009) and the lemmatizer has an estimated accuracy of 99.55% (Ingason et al. 2008).

The Database of Modern Icelandic Inflections (DMII; Beygingarlýsing íslensks nútímamáls, BÍN)

The DMII (Bjarnadóttir 2012a) is a collection of approximately 270,000 Icelandic lemmas and their inflectional forms. In total, the database contains around 5.8 million inflectional forms, each listed along with its corresponding lemma and a grammatical tag identifying which precise inflectional form it is.

The Icelandic Frequency Dictionary (IFD; Íslensk orðtíðnibók) corpus

The IFD corpus (Pind et al. 1991) contains over 500 thousand running words from contemporary literary works of various genres, including fiction, biographies and some informative writings. It is an annotated corpus where each word has been assigned a lemma and a grammatical tag. The annotation has undergone human review and can thus be assumed to have a very low error rate.

The Tagged Icelandic Corpus (Mörkuð íslensk málheild, MÍM)

The Tagged Icelandic Corpus (Helgadóttir et al. 2012) consists of around 25 million running words from a wide variety of contemporary sources, including newspapers, books of various genres and websites. It is annotated and each word has been assigned a lemma and a grammatical tag. However, unlike the IFD corpus, the annotation has not undergone human review and thus the error rate is higher.

Íslenskur orðasjóður

Íslenskur orðasjóður (Hallsteinsdóttir et al. 2007) is a corpus of approximately 250 million running words, sourced from the National and University Library of Iceland's archive of web pages hosted on Icelandic domains (i.e., domains ending with .is). As large quantities of automatically collected web data are inherently noisy, effort has been taken to minimize the presence of foreign language text and ill-formed sentences. The usefulness of this resource for spelling correction may be limited by the fact that it probably contains a large number of misspelled words and ungrammatical sentences. Nevertheless, such a large corpus of modern Icelandic may prove useful for specific purposes.

Annotated compound word list

An extensive list of approximately 270 thousand Icelandic compound words annotated with information on how they are split as well as their word class has been compiled by Bjarnadóttir (2012b). This list, which is an unpublished work in progress, was made available for the purposes of this research.

3 Correction of OCR Text

This chapter introduces methods capable of identifying and correcting nonword and real-word errors. It is split into three parts. In the first part, methods which can be used to identify compound words which are not in the spellchecker's lexicon are introduced. These methods are then adapted to make use of available research data on Icelandic compounds. The second part introduces methods for identifying nonword errors, generating correction candidates for such errors and assigning these candidates a probability of being correct. The methods which are chosen for evaluation are adapted, when deemed appropriate, to better suit the task of correcting Icelandic OCR errors. The third part introduces several different approaches to the correction of real-word errors. These approaches are considered with regard to the challenges presented by Icelandic (i.e., morphological richness, limited language resources and data sparseness). Two such methods, which are further adapted to overcome these challenges, are chosen for evaluation.

3.1 Decompounder

3.1.1 Compound Words in Icelandic

Icelandic is a compounding language, which means that under certain conditions, two words may be combined to form a new word. Words formed in this fashion are known as compounds or compound words. The first part of a compound is called its modifier and the second its head, assuming binary branching (Bjarnadóttir 2005). Modifiers and heads may be of any word class, although some are more productive than others (nouns and adjectives in particular). The rules for compounding are recursive, meaning that modifiers and heads may be compounds themselves.

Most Icelandic compounds fall into one of three categories:

- **Stem compounds**, where the modifier is the stem of a word. Examples of stem compounds include *húsfluga* 'housefly' (*hús* 'house' + *fluga* 'fly') and *lykilorð* 'password' (*lykill* 'key' + *orð* 'word'). Sound changes may occur in the stem, for example in the compounds *raddbönd* 'vocal chords' (*rödd* 'voice' + *band* 'chord') and *mannfræði* 'anthropology' (*maður* 'man' + *fræði* 'studies').
- **Genitive compounds**, where the modifier is in the genitive case, singular or plural. Examples of genitive compounds include *fararstjóri* 'tour guide' (*för* 'tour' + *stjóri* 'leader') and *fatabúð* 'clothing store' (*föt* 'clothes' + *búð* 'store'). Again, sound changes may occur in the modifier.
- **Joined compounds**, where the modifier is affixed with a link morpheme. These morphemes (which are not inflectional endings) may be any of the characters *a*, *i*, *s*, *u* or *(a)n*. Examples of joined compounds include *arðsemismat* 'profitability

analysis' (*arðsemi* 'profitability' + *mat* 'analysis') and *fiskibátur* 'fishing boat' (*fiskur* 'fish' + *bátur* 'boat').

Some compounds fall into other minor categories, including dative compounds, where the modifier is in the dative case, as in *ísilagður* 'ice-covered' (*ís* 'ice' + *lagður* 'covered'), and compounds where the modifier is inflected along with the head, as in *litliputti* 'little finger', acc. *litlaputta* (*lítill* 'little' + *putti* 'finger'). The latter is referred to as internal inflection. However, such compounds are relatively uncommon (Bjarnadóttir 2002).

3.1.2 Decompounding Icelandic Compounds

Decompounding (also known as compound splitting) is the act of breaking a compound word into its constituent parts. In languages where compounds are very productive, it can often prove quite difficult to construct lexicons with a good coverage of their vocabulary. For these languages, decompounding can improve the results of a wide variety of language processing tasks involving lexicons and language models. For example, it has been used with good results in machine translation (Brown 2002), information retrieval (Braschler et al. 2003) and speech recognition (Larson et al. 2000). For the task of spelling correction, decompounding makes it possible to identify compound words which are not present in the lexicon, reducing the number of well-spelled words falsely identified as being nonword errors.

Approaches to decompounding generally fall into one of two categories depending on whether they are based on bilingual or monolingual corpora.

Bilingual Corpora

Decompounding algorithms may be based on bilingual corpora (Brown 2002; Koehn and Knight 2003). A bilingual corpus (also known as a parallel corpus) consists of the same text in two different languages, generally aligned by sentences or words. One of the languages is the target language for which the decompounder will be constructed and the other is one where compounds are very uncommon or nonexistent (e.g., English). For a word-aligned corpus, compound words in the target language can often be identified as they should be aligned to multiple words in the other.

Koehn and Knight (2003) describe an approach which uses a bilingual corpus to identify compound splits where the parts can be translated so that they match up with the translation of the compound itself. For example, considering the Icelandic compound *varaforseti* 'vice president' (*vara* 'vice' + *forseti* 'president'), one can see that when correctly split, the English translation of its compound parts correspond to the translation of the compound itself.

As yet, no bilingual corpus between Icelandic and another language is readily available.

Monolingual Corpora

Other decompounding algorithms rely on monolingual lexicons or text corpora (Monz and de Rijke 2001; Alfonseca et al. 2008). For a potential compound, they may attempt to find all possible splits under the assumption that any word form which is present in the

lexicon or corpus is both a valid modifier and head (optionally allowing for link morphemes between compound parts). Of practical concern with this approach is the fact that there is no guarantee that a compound part exists as an independent word. Since such compound parts, known as bound constituents (Bjarnadóttir 2005) do, by definition, never appear on their own in a text corpus or a lexicon, they would not be recognized by the decompounder. Also problematic is the assumption that any word form appearing in the lexicon or corpus can be either a modifier or a head, which is often not true.

A lexicon of known modifiers and heads can also be used in order to split potential compounds (Schiller 2005; Marek 2006). With such a lexicon, words can be broken down into sequences of modifiers and heads, avoiding the problem of sound changes and bound constituents altogether. This approach also makes it possible to estimate a given compound part's probability of occurring either as a modifier or a head.

Research material on Icelandic compounds contains an extensive list of Icelandic compound word lemmas, along with information on the word class and the position of the split between modifier and head for each lemma (Bjarnadóttir 2012b). This list was made available for the purpose of the research described in this thesis. In the following section, a method of compiling a list of known compound parts from such a list is presented.

3.1.3 Constructing a List of Compound Parts

The compound list contains approximately 270,000 compound lemmas, each containing information on where the split between the modifier and the head lies and annotation of word class.

Table 3.1 An excerpt from the compound list

Compound	Word class
dýptar_lóð	nn
dýptar_mat	nn
dýptar_mæla	v
dýptar_mæling	nf
dýptar_mælir	nm

Table 3.1 is an excerpt from the compound list. Every word in the list is a compound and the underscores mark the separation between modifier and head. Additionally, each compound is annotated with its word class (e.g., *nn* for neuter noun, *nm* for masculine noun, *v* for verb etc.)

Expanding the Compound List

While the compound list is quite extensive, it is still limited in the sense that it only contains the lemmas of compound words. Lemmas alone will not suffice to compile a list of compound parts, as they may be only one of a word's many inflectional forms. For example, while the list contains the compound lemma *dýptar_mæling* 'depth

measurement' (revealing that *dýptar* 'depth' can appear as a modifier and *mæling* 'measurement' as a head), nowhere does it explicitly state that *dýptar_mælingarnar* 'the depth measurements' is a compound word form as well.

As yet, there is no readily available list of inflected compound words in Icelandic annotated with information on their split. However, it is possible to construct such a list using already available resources. The DMII (see section 2.3.4) contains inflectional information on most of the compounds in the list above. This information can be used to expand it to include all of the compounds' inflectional forms and not only their lemmas.

If a compound lemma has an entry in the DMII with a matching word class, all of its inflectional forms are added to the compound list. Each inflectional form will contain the split marker at the same location as the lemma and will be annotated with the same word class. If the compound lemma does not have an entry with a matching word class in the DMII but its head does, the compound is inflected according to that instead. Nothing is added to the list if the DMII lacks inflectional information for the compound or its head.

Table 3.2 Inflections of dýptarmæling *which will be used to expand the compound list*

Compound	**Word class**
dýptar_mæling	nf
dýptar_mælingu	nf
dýptar_mælingar	nf
dýptar_mælinga	nf
⋮	⋮

For example, if the word *dýptarmæling* from the compound list has an entry in the DMII with a matching word class (in this case *nf*, which stands for feminine noun), all of its inflectional forms (*dýptarmælingu*, *dýptarmælingar*, etc.) are added to the list. The split marker is inserted at the same position as in the lemma, right after the modifier *dýptar* (*dýptar_mælingu*, *dýptar_mælingar*, etc.) and each inflectional form is annotated with the same word class. If *dýptarmæling* is not in the DMII but *mæling* is (and, like the compound, is a feminine noun), the compound's inflectional forms are generated by prefixing its modifier (*dýptar*) to all of the head's inflectional forms (*mæling*, *mælingu*, *mælingar*, etc.). This gives identical results.

The expanded list contains over 2.8 million compound word forms. It can be used to estimate the ratio of compounds in the IFD corpus, as the DMII contains every inflectional word which appears in it and the compound list includes every compound in the DMII. The IFD corpus includes 515,727 running words (excluding numbers and punctuation marks), of which 65,748 (12.7%) are present in the expanded compound list. Furthermore, there are 57,678 word forms in the IFD corpus, of which 30,450 (52.8%) are compounds. This analysis shows how prevalent compound words are in Icelandic text.

Determining Compound Part Frequencies

Using the expanded list, it is possible to determine the frequency with which individual compound parts appear. Their frequencies will be compiled from three text corpora: Íslenskur orðasjóður, MÍM and the IFD corpus (see section 2.3.4). These corpora contain around 275 million running words in total. Every time a compound word occurs, the frequency of each of its compound parts is increased by one, either as a modifier or as a head (whichever the case may be).

For example, the word *varaaflgjafi* 'backup power generator' is a compound word which, according to the compound list, is composed of the modifier *vara* 'backup' and the head *aflgjafi* 'power generator'. Each occurrence of the compound in the corpora increases the modifier count of *vara* and the head count of *aflgjafi* by one. However, *aflgjafi* is also a compound, composed of the modifier *afl* 'power' and the head *gjafi* 'generator'. The modifier count of *afl* is also increased by one and so is the head count of *gjafi*. Thus, a single occurrence of this compound results in the frequency of four compound parts being increased by one.

Also counted is how often a compound part appears independently, that is on its own and not within a compound.

Table 3.3 Examples of frequency of words and compound parts in the corpora

Word/Compound part	Word frequency	Modifier frequency	Head frequency
varaaflgjafi	12	-1	1
aflgjafi	114	-1	12
afl	4642	3034	4730
gjafi	52	-1	5898

Table 3.3 shows the word and compound part frequencies of several word forms in the three corpora. Compound parts which do not occur in the lexicon receive a word frequency of -1, which may suggest that they are bound constituents which cannot stand alone. Similarly, a modifier or head frequency of -1 means that the word form or compound part in question did not appear as such in the expanded compound list. This is distinct from a frequency of 0 which means that the word form or compound part is known (i.e., is present in the lexicon or compound list) but did not occur in any of the corpora.

3.1.4 Finding Candidates for Splits

In this section, a known method of splitting potential compounds into known compound parts is introduced. First, every possible binary split of a word is generated. Such a split is considered to be a valid candidate if its first part is a known modifier and the second is a known head, or if either part can be further split into known modifiers and heads. The word itself (with no splits) will also be considered to be a candidate unless it is a known compound word (i.e., it is present in the compound list).

In this work, two assumptions are made in order to limit the number of candidates. The first is that if a known compound part is not a known compound word, then it is a base word (i.e., a non-compound). Since base words should not be split, the only candidate that should be generated for one is the base word itself (with no splits). For example, *bursti* 'brush' is a known head, appearing in compounds such as *tannbursti* 'toothbrush'. Since *bursti* is a known compound part but not a known compound (i.e., does not have an entry in the compound list), it is considered to be a non-compound and its only candidate will be itself. This results in no split being made. On the other hand, *tannbursti* is both known to be a head (appearing in *rafmagnstannbursti* 'electric toothbrush') and a compound word. This means that it will be split as usual, its candidates being any plausible binary split of the word.

The other assumption is that only the candidates which consist of the fewest number of known parts should be considered. Consider the following candidates for the word *virðisaukaskatturinn* 'the value added tax':

Table 3.4 Binary and minimal splits of n known parts for virðisaukaskatturinn

#	Binary split	Minimal n-ary split with n known parts	n
1	virðis + aukaskatturinn	(virðis, aukaskatturinn)	2
2	virðisauk + askatturinn	((virðis, auk), (as, (katt, urinn)))	5
3	virðisauka + skatturinn	(virðisauka, skatturinn)	2
4	virðisaukaskatt + urinn	((virðis, aukaskatt), urinn)	3
5	virðisaukaskatt + urinn	((virðisauka, skatt), urinn)	3

Two of the candidates (#1 and #3) require only a single split until all of its compound parts are known. However, candidates #4 and #5 require two splits and candidate #2 requires four. Since they do not contain the fewest number of known parts, they are discarded in favor of the other candidates.

Thus, the candidates which are generated are all binary splits which consist of the fewest possible number of known parts.

3.1.5 Choosing the Most Likely Candidate

A candidate is chosen based on the probability of its minimal n-ary split of n known parts. Two methods for assigning probability to splits have been considered: a frequency-based method and a probability-based method.

Frequency-Based Method

Koehn and Knight (2003) describe a method which relies on the assumption that the more frequently a word form occurs in a corpus, the more likely it is to appear as a compound part. For a given word w, the n-ary split S with the highest geometric mean of the frequencies of its compound parts p_i is chosen:

$$argmax_S \left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}}$$

There are several disadvantages to this method. One is that it does not distinguish between modifiers and heads, which is problematic because not all compound parts can be both, and those that can might more commonly occur as one than the other. It is therefore possible that a split might be assigned a non-zero probability even if its structure does not conform to the rules on Icelandic compounding (i.e., it contains illegal modifiers or heads). Another drawback lies in the simplifying assumption that compound part frequencies are proportional to word form frequencies, which is not always true (e.g., in the case of bound constituents). These disadvantages can be alleviated by making use of the head and modifier frequency of the compound parts rather than their word frequency. In this work, the frequency based method is adapted to make use of the expanded compound list. What follows is the implementation of this adapted method, which makes use of modifier and head frequencies rather than word frequencies.

For a given word w, the n-ary split S, consisting of modifiers M and heads H, with the highest geometric mean of the frequencies of its modifiers, m_i, and its heads, h_j, is chosen:

$$argmax_S \left(\left(\prod_{m_i \in M} mod\_count(m_i) \right)^{\frac{1}{n_m}} * \left(\prod_{h_j \in H} head\_count(h_j) \right)^{\frac{1}{n_h}} \right)$$

where $n = n_m + n_h$.

For example, consider the two candidate splits for *virðisaukaskatturinn* 'the value added tax': (virðis, aukaskatturinn) and (virðisauka, skatturinn). The known compound parts and their frequencies are:

Table 3.5 Compound part frequencies for the potential parts in virðisaukaskatturinn

Compound part	Modifier count	Head count
virðis	7451	588
virðisauka	6919	-1
skatturinn	-1	413
aukaskatturinn	-1	1

The geometric mean of the compound part frequencies of each split are:

(virðis, aukaskatturinn): $7.451 * 1 = 7.451$

(virðisauka, skatturinn): $6.919 * 413 = 2.857.547$

Thus, (virðisauka, skatturinn) is correctly chosen as the most likely split.

Probability-Based Method

Schiller (2005) assigns each compound part a weight depending on its probability of appearing in a corpus. A compound part p is assigned the weight

$$W(p) = -\log\left(\frac{count(p)}{N}\right)$$

where $count(p)$ is its word frequency in a corpus containing N running words. The split S with the lowest sum of its compound part weights is chosen as the most likely split:

$$argmin_S \sum_{p_i \in S} W(p_i)$$

This approach shares some of the disadvantages of the frequency-based method described by Koehn and Knight due to its reliance on word form frequency. However, these disadvantages can be alleviated by adapting the method to model the weight of heads and modifiers independently. The weights can then be defined as

$$W_{mod}(m) = -\log\left(\frac{mod\_count(m)}{N}\right) \qquad W_{head}(h) = -\log\left(\frac{head\_count(h)}{M}\right)$$

where N and M are the total number of modifiers and heads (respectively) observed in some corpus. The likeliest split would then be:

$$argmin_S\left(\sum_{m_i \in M} W_{mod}(m_i) + \sum_{h_i \in H} W_{head}(h_i)\right)$$

Applying this method on the candidate splits for *virðisaukaskatturinn* (where $N = 41{,}208{,}872$ and $M = 41{,}395{,}720$) gives the following results:

(virðis, aukaskatturinn): $-\log\left(\frac{7451}{41208872}\right) - \log\left(\frac{1}{41395720}\right) = 26.16$

(virðisauka, skatturinn): $-\log\left(\frac{6919}{41208872}\right) - \log\left(\frac{413}{41395720}\right) = 20.21$

Like in the frequency-based method, (virðisauka, skatturinn) is correctly chosen as the most likely split.

A Comparison of the Frequency-Based and Probability-Based Methods

In order to relate the frequency based method of Koehn and Knight (KK) to the probability based method of Shiller (S) it is useful to recast the former method into a probability setting, using the fact that:

$$argmax_S \left(\left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}} \right)$$

$$= argmax_S \left(\log \left(\left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}} \right) - \log(N) \right)$$

since log is an increasing function and $\log(N)$ is a constant

$$= argmax_S \left(\frac{1}{n} \sum_{p_i \in S} \log\big(count(p_i)\big) - \log(N) \right)$$

$$= argmin_S \left(\frac{1}{n} \sum_{p_i \in S} W(p_i) \right)$$

where $W(p_i) = -\log\left(\frac{count(p_i)}{N}\right)$

Thus, the criteria for choosing the most likely split in KK on one hand and S on the other are identical apart from the factor $1/n$ and the fact that in S we may replace N by M or H. The effect of the former change is that the weight of larger compounds is somewhat reduced in KK compared with S. The effect of the latter change is that if M is larger than H then the weights of the modifiers are reduced as compared with the weights of the heads and vice versa.

Due to the similarity of the two approaches, only the frequency-based method will be evaluated.

3.2 Nonword Error Correction

In this work, a nonword is considered to be any word which is not generally accepted to be part of the vocabulary of the language it is written in. For example, *fleirri* is a nonword with regard to Icelandic, as there is no word in the language's vocabulary which is written that way. Should *fleirri* be mistakenly written instead of *fleiri* 'more', it is considered to be a nonword error. However, if it is intentionally written as an example of a common misspelling in Icelandic, it is not a nonword error as it is not meant to be corrected (although it is still a nonword). As there is no easy way to separate nonwords written as intended from nonword errors, any nonword will be regarded as a potential error. In a real-world application, users would get the chance to ignore nonwords which are not errors or to add them to the spellchecker's lexicon.

This section introduces known methods to identify nonwords, to find likely candidates for their correction and to rank the candidate corrections depending on their probability of being correct. When appropriate, these methods are adapted to overcome some of the challenges presented by correcting Icelandic OCR text.

3.2.1 Detecting Nonword Errors

A lexicon based approach is used to detect nonword errors. The lexicon is primarily based on the DMII, which contains roughly 2.8 million distinct Icelandic word forms. The word forms from the DMII are complemented by a list of approximately 2,900 uninflectable words acquired from the Árni Magnússon Institute for Icelandic Studies. While the DMII may possibly be the largest single collection of well-spelled Icelandic words, it is still by no means complete with regard to coverage of the vocabulary. This is in large part due to the rich morphology of the language, especially because of how productive it is in the formation of new compounds. Due to this reason, words which are not in the lexicon will only be classified as nonword errors if they cannot be decompounded.

Thus, in the proposed nonword error detection algorithm, any out-of-lexicon word form which cannot be split by the decompounder will be regarded as a nonword error.

3.2.2 Generating Candidates for Correction

There are many approaches of finding potential corrections of nonword errors. The most studied techniques are based on algorithms which find the minimum edit distance between a nonword and a well-spelled word from a lexicon. The minimum edit distance between two strings is the smallest number of edit operations required to transform one string into the other. There are many different metrics for measuring edit distance, differing in how edit operations are defined and whether or not additional aspects of the strings are considered.

One such metric is the Levenshtein distance (Levenshtein 1966), which counts every deletion or insertion of a single character and every substitution of one character with another as a single edit operation. This definition is something which can be expanded upon, as is done in the Damerau-Levenshtein distance (Damerau 1964) where the transposition of two adjacent characters is also considered a single operation. Additionally, the edit distance can be weighted by other features of the strings such as the phonetic similarities between them (Veronis 1988). While such considerations may prove helpful in the correction of typographical errors, they do not seem to have any obvious benefit over the simpler Levenshtein distance in the setting of OCR post-correction.

Studies on English typographical errors have shown that a large majority of misspelled words contain only a single error. The ratio of single-error misspellings ranges from 80% (Damerau 1964) to 94% (Pollock and Zamora 1984). Using a minimum edit distance algorithm, it is possible to find all well-spelled words in a lexicon which are a single edit operation away from a particular misspelled word. Due to the very high relative frequency of single-error misspellings, the correct word is likely to be among the

candidates (assuming it is present in the lexicon). The question then arises whether or not the same holds true for OCR errors.

Table 3.6 The Levenshtein distance in a portion of the Alþingi digitization project

Edit distance	Error count	Ratio
1	913,376	89.4%
2	93,751	9.2%
3	10,688	1.0%
4+	3,792	0.4%
	1,021,607	**100.0%**

Table 3.6 shows the Levenshtein distance between 1,021,607 misrecognized words and the corrections made to them by proofreaders in Alþingi's digitization project. It reveals that 89.4% of the errors required only a single character to be changed and that 98.6% required two or fewer changes. The ratio of single-error words among misrecognized words in OCR generated text seems to be consistent with that of single-error typographical errors in typed English text.

Furthermore, an examination of words with multi-character errors shows that most are occurrences of a single character being mistaken for two or vice versa (e.g., *k* for *lc* or *in* for *m*). Expanding the Levenshtein's definition of an edit operation to include the correction of any multi-character errors which occurred at least 10 times in the Alþingi digitization project as a single operation yields the following results:

Table 3.7 The edit distance with additional edit operations considered

Edit distance	Error count	Ratio
1	973,693	95.3%
2	37,454	3.7%
3	7,148	0.7%
4+	3,312	0.3%
	1,021,607	**100.0%**

Using this expanded definition, 95.3% of the errors are only one edit operation away from their correction. The ratio of misrecognized words requiring two edit operations or less to fix is increased to 99.0%. This demonstrates that in the vast majority of cases, it should be enough to look for correction candidates which are a single edit operation away from the misrecognized word.

The correction candidates which are generated for a given nonword are all word forms in the lexicon which are a single edit operation away from the nonword, or failing that, two edit operations away. Here, an edit operation is the deletion, replacement or insertion of

a single character or the correction of any multi-character error which occurs at least 10 times in Alþingi's digitization project (as determined from corrections made by proofreaders).

If no candidates can be found in the lexicon, the decompounder will suggest any word form a single edit operation away from the error, which can be split into known compound parts. Failing that, no candidates will be generated.

It can be determined from the analysis on edit distances above that the correct word will be among the candidates in at least 95.3-99.0% of all cases (assuming it is present in the lexicon), which is satisfactory for the task at hand.

3.2.3 Ranking Candidates

Church and Gale (1991) introduce a method to assign probabilities to candidates for the correction of single-error misspellings (assuming the edit operations of the Damerau-Levenshtein distance). Their method is based on a noisy-channel model, which assumes that words are signals being sent through a channel, and that misspellings occur due to noise in that channel. For typographical errors, the channel could be the keyboard or the person typing the word. For OCR errors, the channel is the OCR process itself. The task is then to build a probabilistic model of the channel which can estimate the probability of a candidate correction c being correct given a noisy (i.e., misspelled) word n, $P(c|n)$.

Using Bayes' rule and dropping the constant denominator, the probability can be calculated as $P(c)P(n|c)$. The prior (or language model), $P(c)$, is the probability of appearance of the word c, which can be estimated from a large corpus. The channel model, $P(n|c)$, is the probability of c having been transformed into n by the noisy channel. It is estimated from the likelihood of the transformation (i.e., the single edit operation) from c to n occurring in a training set. For example, $P(absorbent|absorbant)$ is calculated as

$$\frac{words(absorbent)}{N} * \frac{sub(a,e)}{chars(a)}$$

where $words(absorbent)$ is the frequency of *absorbent* in some corpus containing N running words, $sub(a,e)$ is the number of misspelled words where e has been misspelled as a in a training set and $chars(a)$ is the character frequency of a in the training set.

Church and Gale trained their model on the 1988 Associated Press corpus, which contains 44 million running words. To estimate the likelihood of a specific transformation occurring, they made a list of all nonwords appearing in the corpus for which only a single correction candidate could be generated. The nonword was then considered to be a misspelling of the single correction candidate, and the transformation from correct word to the misspelling was counted.

Brill and Moore (2000) improve on this method in several ways. Perhaps most significantly, they expand the set of possible edit operations beyond that which is allowed

by the Damerau-Levenshtein distance. Those operations are complemented by additional transformations learned from a training set of known spelling errors and their corrections. Each transformation learned in this manner will be considered to be a single edit operation.

In this work, the difference algorithm described in (Myers 1986) is used to extract minimal transformations from the training set. For example, it finds that to correct *efnaliagur* to *efnahagur* 'economy', the transformation $li \rightarrow h$ is required. This differs somewhat from the approach used by Brill and Moore, who consider combinations of multiple transformations (e.g., $l \rightarrow h$ and $i \rightarrow \varepsilon$, or $l \rightarrow \varepsilon$ and $i \rightarrow h$) and also the context of error (e.g., $ali \rightarrow ah$). Thus, a much larger number of possible transformations are acquired from each word pair. They explain that this is helpful in cases such as when *antechamber* and *antecedent* are misspelled as antichamber or anticedent, because the real error being made is that *ante* is being confused with *anti* rather than simply *e* for *i*. There is no clear parallel between errors such as these and OCR errors, and for that reason a difference algorithm is used to obtain a minimal number of transformations.

To compute the channel model probability for a given correction candidate of some misspelled word, the characters of the two words are first aligned with one another. It is then calculated as the product of the probability of each transformation from characters in the correction candidate to the corresponding characters in the misspelled word. Unchanged characters are considered to be transformations as well. For example, the correction candidate *ríkið* 'the state' for the OCR error *rílciö* would be assigned the probability $P(r|r) * P(\text{í}|\text{í}) * P(lc|k) * P(i|i) * P(\text{ö}|\text{ð})$. Here, $P(r|r)$ is the probability that the character *r* will be correctly recognized and output as *r*. Similarly, $P(lc|k)$ is the probability that *k* will be misrecognized and output as *lc*.

This method is evaluated on nonword errors from the Alþingi digitization project which were corrected by proofreaders. The testing set can be easily derived from the digitized text by comparing each proofread speech to its unproofread counterpart.

Brill and Moore consider several additional improvements, such as utilizing language models and the positional information of transformations (i.e., whether they occur at the beginning, in the middle or at the end of a word). A language model is not implemented in this work, for the reasons discussed in section 2.3.2. The usefulness of positional information in the context of OCR post-correction is not investigated in this thesis, but may be considered in future work.

An overview of the nonword error correction algorithm is given in figure 3.1.

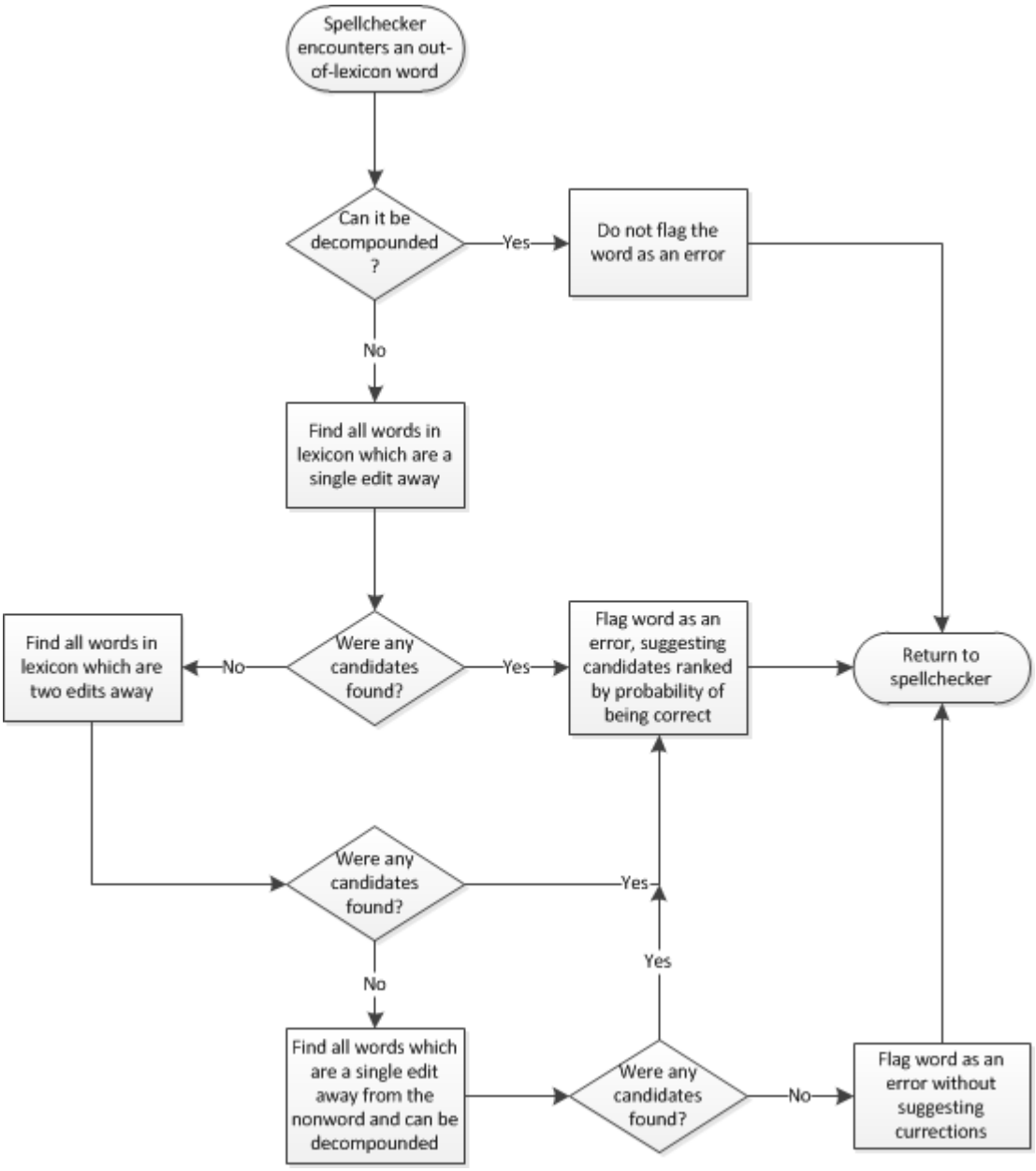

Figure 3.1 Flowchart giving an overview of the nonword correction algorithm

3.3 Real-word Error Correction

In this chapter, several methods for correcting real-word errors will be introduced. These methods fall into various categories, such as statistical-based methods and rule-based methods, though they all tend to share some fundamental qualities in terms of the approach taken. Two methods which appear to be suitable for the task of correcting real-word errors in digitized Icelandic text will be reviewed in further detail.

3.3.1 Detecting and Correcting Real-word Errors

Disambiguation Methods

Word-sense disambiguation is a common task in NLP applications. It is the task of determining which sense (i.e., meaning) an ambiguous word form takes given the context in which it appears. For example, the word *orange* is ambiguous as it may either refer to the fruit or the color. In the sentence "I had eggs, toast and orange juice for breakfast", it is clear from the context that it refers to the fruit. While English speakers would likely have no difficulty discerning its meaning, exactly replicating the process by which they do so is a difficult if not impossible task by the use of software. However, one could make the observation that when the word *orange* is followed by the word *juice,* it very likely refers to the fruit. The words *eggs* and *toast* might also tend to show up more often in the context of *orange* as a fruit than *orange* as a color. Furthermore, a good POS (part-of-speech) tagger would likely tag the word as a noun, effectively ruling out the possibility that it could be referring to a color (adjective). A word-sense disambiguation tool can be trained to recognize which semantic and grammatical properties occurring in the context of an ambiguous word (known as its features) have a stronger connection to one meaning over others, and to choose the most likely meaning on the basis of that knowledge.

Certain similarities can be drawn between word-sense disambiguation and real-word error correction. An example of a common English misspelling is when *where* is confused with *wear* or vice versa. The two words have different meanings and tend to appear in different semantic and grammatical contexts, much like the different meanings of the word *orange*. To determine whether one word has been confused with the other, one could apply word-sense disambiguation to find out whether the context better fits the meaning (and therefore also the spelling) of *wear* over that of *where*. In this way, real-word error correction can be cast as a disambiguation problem.

Words which are commonly confused with one another are placed together in what are called confusion sets. For example, one could define the confusion sets {*wear*, *where*} and {*their*, *they're*, *there*}. Every time a word belonging to some confusion set (i.e., a confusion word) occurs in a text document, a spellchecker could evaluate which word from the same set has the best fit to the context in which it appears. If the word with the best fit is not the one which was written, it can be flagged as an error and the most likely alternative from the same confusion set suggested as a correction.

Many different disambiguation methods have been applied to the task of real-word spelling correction with good results, including a Bayesian-based approach (Golding 1995)

and a Winnow-based approach combined with a weighted-majority algorithm (Golding and Roth 1999). Before these methods are described in more detail, some alternative approaches will be reviewed.

Rule-based Methods

One way of correcting real-word errors is by the use of rules. For example, a rule might state that *wear* should be changed to *where* if it is immediately followed by the word *are*. Compiling a list of such rules by hand is a time consuming process which is infeasible on a large scale.

Mangu and Brill (1997) describe a transformation-based learning approach for automatically generating rules for spelling correction. Confusion sets are defined as above and the features of each confusion word (grammatical and semantic properties observed in their context) are extracted from a large text corpus. For example, one confusion set might be {*wear*, *where*} and the features of *where* might include the word *going* occurring within ± 3 words of it or it being followed immediately by the word *is*. From these features, the potentially useful rules "change *wear* to *where* if *going* occurs within ± 3 words of it" and "change *wear* to *where* if it is followed immediately by *is*" would be automatically constructed. Rules are created in this manner from all the features of every confusion word.

Initially, the spellchecker does not make use of any of the rules but instead relies solely on a baseline predictor. The baseline predictor may simply assume that the most common word in a confusion set is always correct (regardless of context), but it could also be something more sophisticated like a full-fledged spellchecker. The prediction accuracy may then be improved in a training phase where the rules which prove to be most useful are incorporated into the spellchecker.

The training takes place for one confusion set at a time. First, the baseline predictor will be passed a set of training examples consisting of sentences which contain words from the current confusion set. Every time a confusion word occurs in an example, the baseline predictor will attempt to determine which word from the confusion set is most likely to be correct. The predictor's accuracy is evaluated by an objective function. Assuming that the training data does not contain spelling errors, it would suffice to compare the predictor's guesses to the original sentences.

Next, the rules for the current confusion set are evaluated by applying them to the baseline predictor's guesses. A rule will take precedence over the predictor's guess should they disagree. The rule which improves the predictor's accuracy by the greatest amount is adopted by the spellchecker. From then on, the spellchecker will first pass every example to the baseline predictor and will then apply the rule to the example, possibly altering the predictor's guess. The process is then repeated by evaluating every remaining rule on the output of the improved spellchecker. Again, the rule which results in the greatest improvement to the prediction accuracy will be adopted by the spellchecker and will be applied directly after the previously adopted rule from then on. This process continues until no remaining rule can improve the prediction accuracy, at which point the training ends. This is repeated for every confusion set.

Mangu and Brill (1997) find that when evaluated on the Brown Corpus (Kucera and Francis 1967), this method has an average prediction accuracy of 93.15% for English. Their experiments showed that when evaluated on the same corpus, it performed better than a Bayesian classifier (Golding 1995), which had an average accuracy of 91.15%, but it was in turn outperformed by the Winnow classifier (Golding and Roth 1999), with an average accuracy of 94.01%.

Statistical Language Model Methods

A statistical language model (SLM) may be used to detect and correct real-word errors (Jurafsky and Martin 2009). A SLM can estimate the probability of a sequence of n words (n-grams) appearing in a sentence based on how often it occurs within some large text corpus. Real-word errors are indicated by the presence of n-grams which are not known to the language model or where the probability of appearance is very low. Corrections might be known n-grams which are a short edit distance away from the suspected error.

A considerable benefit of this approach is that it does not require potential real-word errors to be manually defined beforehand in a confusion set. However, a large amount of text is required in order to construct an accurate language model. This approach might therefore not be feasible for morphologically rich languages where language resources (e.g., text corpora) are limited.

3.3.2 Real-word Errors in OCR Text

An analysis of the errors which were corrected during the proofreading phase of Alþingi's digitization project reveals that 94% were nonword errors and 6% were real-word errors. However, the ratio of real-word errors may be underestimated due to the use of spellchecking software which only detects nonword errors. A human proofreader with no assistance from a spellchecker is unlikely to find more than 70-75% of the errors present in a document displayed on a computer screen (Creed et al. 1988). This implies that a higher portion of real-word errors than nonword errors went unnoticed by the proofreaders, and that the actual ratio of real-word errors present in the OCR text is likely to be somewhat higher than 6%.

The analysis also showed that some words were far more likely than others to result in a real-word error when misrecognized by the OCR software. In fact, only 10% of the words where misrecognition resulted in a real-word error accounted for 89% of all real-word errors. These results suggest that the flexibility afforded by SLM-based methods (i.e., theoretically being able to detect any real-word error without it being defined beforehand) may not be necessary in order to have a chance of correcting the vast majority of these errors. Additionally, it remains an unanswered question whether the amount and quality of available language resources for Icelandic suffice for the construction of an SLM which is well suited to the task of context-sensitive spelling correction.

For the above reasons, only methods based on confusion sets will be considered. The Winnow-based approach (Golding and Roth 1999) will be evaluated as Mangu and Brill (1997) found that it outperformed both the rule-based approach which they described as

well as the Bayesian-based approach (Golding 1995) mentioned above. Due to its simplicity and its similarity to the Winnow-based approach, the Bayesian-based approach will also be evaluated for comparison purposes. While the rule-based approach may be no less suitable for the task at hand, its evaluation will be left to future work as the purpose of the evaluation is not to determine which confusion set based method gives the best results, but whether they are viable at all for correcting real-word errors in digitized Icelandic documents. An evaluation of the Winnow and Bayesian-based approaches is sufficient for that purpose.

3.3.3 Naive Bayesian Classifier

Golding (1995) describes an approach to real-word spelling correction using a Bayesian classifier. A classifier is an algorithm which takes a set of features representing some object and uses them to determine which class this set belongs to (White 2000). In the context of spelling correction, classes might be words which are commonly confused with one another, e.g., *weather* and *whether*. An object might be such a word (referred to as a target word), along with the sentence in which it occurs. The features could be some semantic or grammatical qualities of the target word which could help determine which word is correct (i.e., which class they belong to).

Features

Golding (1995) used two types of features: context words (also known as co-occurrences) and collocations.

Context words are word forms which occur within $\pm k$ words of a target word. For example, consider the sentence "The _ in summer is warm and sunny." where the underscore represents a target word which could either be *weather* or *whether*. For $k = 6$, the context words of the target word are *the*, *in*, *summer*, *is*, *warm*, *and* and *sunny*. The presence of the words *summer*, *warm* and *sunny* in the target word's context strongly implies that it should be *weather*.

Collocations are patterns of up to l contiguous word forms and POS tags[1] in which the target word occurs. Among the collocations of the target word in the sentence above (for $l = 3$) would be "_ in summer" and "DT _", where the POS tag DT stands for determiner. Both of these collocations would imply that the target word should be *weather*.

In this work, the collocations will be expanded to include lemmas of words appearing in the collocation patterns. This is done in an attempt to mitigate some of the difficulties presented by a highly inflectional language. Consider the pair of Icelandic words *bil* 'space' and *bíl* 'car' which OCR software might have difficulty telling apart. The verb *keyra* 'drive' has 29 unique inflectional forms, most of which can directly precede the word *bíl*. Using lemmas in the collocation patterns (in addition to word forms and POS tags) makes it possible to combine every single collocation where the word *bíl* is preceded by an inflectional form of *keyra* under a single feature. This means that the collocation patterns

[1] In this work, the IceNLP toolkit (see section 2.3.4) is used to tag text.

“keyrði$_{cw}$ _”, “keyrir$_{cw}$ _” and “keyrðum$_{cw}$ _” could be combined under the feature “keyra$_{cl}$ _” (where *cw* stands for context word and *cl* stands for context lemma).

For the same reason, context lemmas will also be used as features. A context lemma is the lemma of a context word. This makes it possible to combine every occurrence of some inflectional form of *keyra* in the context of *bíl* to be combined under a single feature.

Training

The classifier is trained on a large text corpus to learn how strongly each feature indicates that a given confusion word is correct. Training the Bayesian classifier involves going through the corpus and counting every occurrence of a confusion word as well as the features which are present in its context.

Golding (1995) only trains the classifier on features which have significant correlation with their target word in an effort to reduce the amount of noise and other irrelevant attributes among the features. Golding and Roth (1999) later demonstrated that pruning the feature set in this manner reduces classification accuracy. They conjecture that this is due to the fact that it filters out many “small disjuncts”, features which cannot be correlated with the confusion set as a whole but may nonetheless prove helpful when classifying a small portion of the training set. They find that pruning features which only occur once in the training set yields considerably better results than more aggressive pruning, as in Golding (1995). For this reason, the only pruning that will take place in this work is on features which only occur once in the training set. Though this pruning is minimal, it roughly cuts the number of features extracted from the training corpus by half.

Classification

It is the classifier’s task to determine which word in a confusion set $C = \{w_1, w_2, \dots, w_n\}$ is correct, given a set $\mathcal{F}$ of active features which have been observed in its context. Bayes’ rule with the conditional independence assumption is used to calculate $P(w_i|\mathcal{F})$, the probability of $w_i \in C$ being correct given the presence of features $\mathcal{F}$:

$$P(w_i|\mathcal{F}) = \left(\prod_{f \in \mathcal{F}} P(f|w_i) \right) \frac{P(w_i)}{P(\mathcal{F})}$$

The word with the highest probability is then chosen as the correct one. In this work, features which were not observed in the training set are assigned a probability of $0.01/m$, where m is the frequency of w_i. This is known as no-matches-0.01 smoothing (Kohavi et al. 1997).

3.3.4 Winnow Classifier

The Winnow algorithm is a classifier which is known for its good performance in settings where the feature space is very large while only a few features are relevant (Littlestone

1988). The algorithm has been shown to give good results when applied to the task of context-sensitive spelling correction (Golding and Roth 1999).

Training

A Winnow classifier is created for every word in a confusion set. The classifiers are trained on example sentences containing at least one word from their confusion set. For every occurrence of a confusion word in the examples, each classifier must attempt to determine whether its assigned word could be correct in the same context.

Each classifier could be considered to contain a network composed of all the features occurring in the training set. When a classifier observes some feature f for the first time during the training phase, an arc with a default weight of 0.1 is drawn from f to the classifier. If the total weight of the active features which are connected to the classifier is greater than some threshold θ, the word is considered to be well-spelled and the classifier returns a classification of 1. Otherwise, it returns a classification of 0.

Weights are updated only when a classifier makes an incorrect prediction. If the classifier predicts that its word, w, is incorrect when it is correct (i.e., the example is positive), the weight on the arc connecting each active feature $f \in \mathcal{F}$ is increased:

$$\forall f \in \mathcal{F}, w_f \leftarrow \alpha \cdot w_f$$

where w_f is the weight on the arc connecting f to the classifier and $\alpha > 1$ is a promotion parameter. This reduces the likelihood of the mistake being repeated if the word is seen again in a similar context. If the classifier predicts that its word is correct when it is not (i.e., the example is negative), the weights on the arcs are demoted:

$$\forall f \in \mathcal{F}, w_f \leftarrow \beta \cdot w_f$$

where $0 < \beta < 1$ is a demotion parameter. Golding and Roth (1999) used the threshold $\theta = 1$, the promotion parameter $\alpha = 1.5$ and varied the demotion parameter β from 0.5 to 0.9 (see discussion on the weighted-majority algorithm below). In this work, the same values will be used, except for β, which will be varied from 0.1 to 0.9 for reasons explained below.

Weighted-Majority Voting

Golding and Roth (1999) make the observation that the less overlap there is between the usages of words in a confusion set, the more appropriate it becomes to use a low demotion parameter (i.e., demote the weights considerably). For example, the words *weather* and *whether* can rarely be replaced with one another. Thus, in the case of {*weather*, *whether*}, a low demotion parameter is appropriate because the classifier is very likely making a mistake when its classification disagrees with the training example. On the other hand, the same does not apply to the confusion set {*between*, *among*}, as the two words can often (but not always) be replaced with one another. In that case, a higher demotion parameter may be appropriate as the classifier is not necessarily making a mistake if its classification disagrees with the training example.

To address this issue, Golding and Roth (1999) use five classifiers for each confusion set (forming what they call a cloud of classifiers), each differing in the value of their demotion parameter (0.5, 0.6, 0.7, 0.8 and 0.9). A weighted-majority (WM) algorithm is used to combine the classifiers, weighing each one by the accuracy of their predictions (Littlestone and Warmuth 1994). The *j*th classifier receives the weight γ^{m_j}, where γ is a constant such that $0 < \gamma < 1$ and m_j is the number of mistakes it has made thus far. Golding and Roth (1999) used an initial value of $\gamma = 1.0$, decreasing with time as more training examples are seen. This was done in order to avoid weighing mistakes of the initial hypothesis too heavily. In this work, a constant value of $\gamma = 0.99$ is used. Additionally, only the last 100 mistakes of a classifier are tracked by m_j. The prediction value of the WM algorithm is

$$\frac{\sum_j \gamma^{m_j} C_j}{\sum_j \gamma^{m_j}}$$

where C_j is the classification of the *j*th classifier (1 if the sum of its weight is 1.0 or greater, else 0). The confusion word with the highest prediction value is assumed to be correct. If the predictions of two or more confusion words are equal (e.g., if every classifier of each confusion word returns a classification of 1, resulting in a total prediction of 1 for both words), the choice between them is arbitrary.

In this work, the number of classifiers in each cloud is increased to 9, with β ranging from 0.1 to 0.9. This is done because many of the real-word errors in the Alþingi documents appear in very disjoint contexts (e.g., *liður* 'item' and *líður* 'feels'). In such cases, a very low demotion parameter may yield better results.

One potential issue with the WM algorithm used by Golding and Roth (1999) is the fact that the individual classifications are binary values. There is no distinction made between a classifier with a total feature weight of 0.09 and 0.99 (which would both return a classification of 0), or of 1.01 and 2.01 (both resulting in a classification of 1). This may be problematic in situations where two or more confusion words are tied for the highest prediction, resulting in an arbitrary choice between them. If the total feature weight of each classifier of the former classifier cloud were 1.01, 1.13, 1.09, 1.07 and 1.05 and the weights of the latter cloud were 1.79, 1.92, 2.11, 2.04 and 2.05, one would intuitively wish to select the confusion word of the latter cloud. For this reason, a more generalized version of the WM algorithm is used in this work

$$\frac{\sum_j \gamma^{m_j} x_j}{\sum_j \gamma^{m_j}}$$

where x_j is the total feature weight of the *j*th classifier.

A general overview of the real-word error correction algorithm is given in figure 3.2.

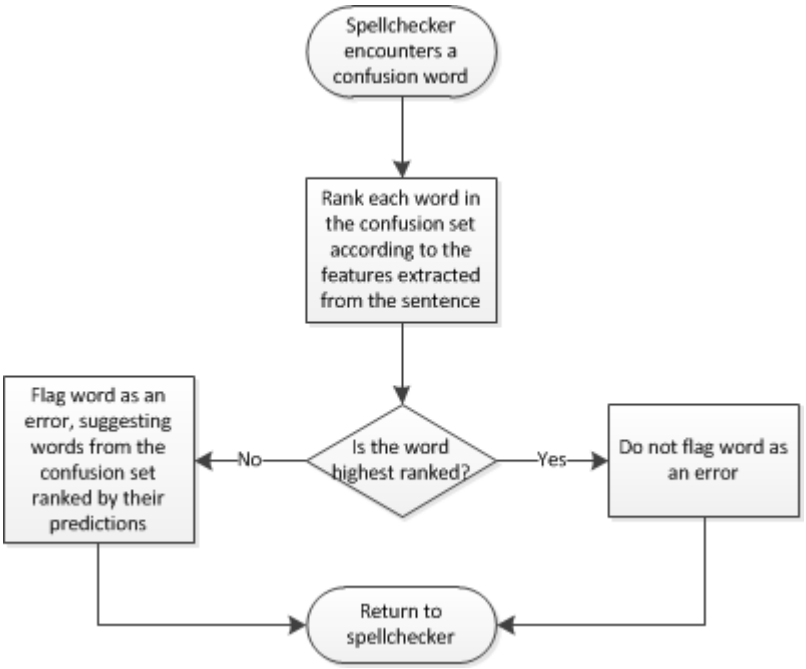

Figure 3.2 Flowchart giving an overview of the real-word error correction algorithm

4 Evaluation

In this chapter, the spellchecker's ability to detect and correct nonword and real-word errors is evaluated.

4.1 Nonword Errors

Proofreaders corrected 959,052 nonword errors which had been introduced during the digitization of the parliamentary speeches. In total, this accounts for 93.88% of all corrections made by proofreaders. In this section, the decompounder's performance is evaluated, particularly with regard to its effectiveness when applied during nonword error detection. The correction accuracy of the spellchecker is also evaluated.

4.1.1 Evaluation Settings

In this evaluation, a compound is considered to be any word form which has a meaning, a structure which conforms to the rules on Icelandic compounding and which is generally accepted to be a well-spelled word[1]. A non-compound is any word form which does not meet all three criteria.

The purpose of the decompounder is to identify out-of-lexicon compounds, and thus reduce the number of words mistakenly flagged as nonword errors. Its performance is evaluated in terms of precision and recall. A high precision means that few non-compounds will be mistakenly identified as compound words (reducing the risk that nonword errors will go undetected). A high recall means that most of the compounds will be correctly identified as such.

The decompounder is evaluated on two test sets: a set of 3,994 word forms (of which 1,688 are compounds) acquired from Icelandic Wikipedia articles, and a random sample of 2,000 out-of-lexicon word forms (of which 676 are compounds) from the Alþingi digitization project. In both cases, each word form has been tagged with information on whether it is a compound or not. Compound parts which are one or two characters in length and have a frequency of less than 1,000 are pruned. This is done because there are many such short and uncommon compound parts (e.g., appearing only within one or two proper names, which are likely already in the lexicon), greatly increasing the possibility of a valid split being found for non-compounds. Examples of such compound parts include modifier *hú* from the proper name *Húbert* (with a modifier frequency of 52) and the head *íð* 'craft' from the words *handíð* 'handicraft' and *myndíð* 'visual crafts' (with a head frequency of 8).

[1] Strictly speaking, a compound is any word form which conforms to the rules of compounding.

The spellchecker's ability to detect nonword errors is evaluated by the use of a lexicon, with or without the aid of a decompounder. Its performance is evaluated in terms of precision and recall. A high precision means that few well-spelled word forms will mistakenly be flagged as nonword errors. A high recall means that most nonword errors will be flagged as such. The evaluation is performed on the same test set of 2,000 out-of-lexicon word forms as above.

Lastly, the spellchecker's ability to correct nonword errors is evaluated. The evaluation is performed on all nonword errors which were corrected during the proofreading phase of the Alþingi digitization project.

4.1.2 Detection of Nonword Errors

When evaluated on the test set of word forms from the Wikipedia articles, the decompounder identifies compound word forms with a precision of 99.41% and a recall of 99.41%. These results show that it performs extremely well on text which is both well-formed and well-spelled. However, in this work, its use is limited to out-of-lexicon words only, many of which are misspellings, which may negatively affect its performance.

Evaluated on the Alþingi test set of out-of-lexicon word forms, the decompounder identifies compounds with a precision of 81.13% and a recall of 97.34%. The significant drop in precision means that almost one in every five words forms identified as a compound is in fact a non-compound, likely a nonword error which then goes undetected. The lower the precision, the more nonword errors remain undetected. On the other hand, the lower the recall, the more time users spend having to manually ignore errors or add words to the lexicon. Whether this is a satisfactory balance between precision and recall depends on the importance placed on the correctness of the text once its proofreading has finished. In this case, it is skewed heavily towards recall, sacrificing correctness for speed.

Many of the non-compounds were mistakenly classified because the decompounder managed to piece together sequences of short, and often infrequent, compound parts. For example, the nonword *ríkísstjórn*, a misrecognition of *ríkisstjórn* 'state government' (*ríkis* 'state', *stjórn* 'government') is split as (*rík* 'rich', (*ís* 'ice', *stjórn* 'government')) and *bensinkaup*, a misrecognition of *bensínkaup* 'gasoline purchase' (*bensín* 'gasoline', *kaup* 'purchase') was split as ((*ben* 'wound', *sin* 'sinew'), *kaup* 'purchase').

In an effort to increase the decompounder's precision, the evaluation is repeated with more aggressive pruning applied. All compound parts which are four letters or less in length and have a frequency of less than 1,000 are removed. Additionally, the compound parts *í*, *á* and *ís* are removed due to the exceptionally high frequency of their occurrence in splits generated for non-compounds. Finally, a word form is only classified as a compound if no word form in the lexicon is a single edit operation away from it (recalling that over 95% of all nonword errors in the Alþingi digitization project are a single edit operation away from their correction, as shown in section 3.2.2). With these additional constraints, the decompounder achieves a precision of 97.37% and a recall of 82.10%.

Table 4.1 The results of the evaluation on nonword error detection accuracy

	Precision	Recall
Lexicon	65.70%	100.00%
Lexicon and decompounder (minimal pruning)	97.73%	88.43%
Lexicon and decompounder (aggressive pruning)	90.84%	98.86%

Table 4.1 shows the results of using a lexicon (as described in section 3.2.1) and the combination of a lexicon and decompounder to detect nonword errors in the Alþingi test set. It is possible to find all nonword errors using only a lexicon. However, this results in a large number of false positives (i.e., low precision), as over one in every three out-of-lexicon word forms is a well-spelled compound. A considerable portion of the user's time will be spent dealing with words mistakenly flagged as errors. When used in conjunction with the decompounder, false positives can be reduced to around one flagged word form in eleven (with aggressive pruning) to around one in every thirty (with minimal pruning). The low number of false positives that occur with minimal pruning come at a cost, as 11.57% of all nonword errors are missed. This cost can largely be mitigated by the use of aggressive pruning, reducing the number of missed nonword errors to only 1.14%.

4.1.3 Correction of Nonword Errors

Table 4.2 shows the results of the evaluation, using aggressive pruning.

Table 4.2 The results of the evaluation on nonword error correction accuracy

Error Category	Total errors	Detected	Corrected
Nonword errors	959,052	99.72%	92.90%
Real-word errors	62,555	0.00%	0.00%
Total	**1,021,607**	**93.62%**	**87.21%**

Using a lexicon in conjunction with a decompounder, it is possible to detect 99.72% of all nonword errors and 93.62% of all errors in the parliamentary speeches which were corrected by proofreaders. Of the nonword errors which are detected, the highest ranked candidate is correct in 93.16% of all cases, meaning that a total of 92.90% of the nonword errors were corrected (accounting for missed errors). In turn, this means that 87.21% of all errors in the parliamentary speeches were corrected.

4.2 Real-word Errors

There were 62,555 real-word errors corrected during the proofreading phase of Alþingi's digitization project, accounting for 6.12% of all corrections made by proofreaders. In this section, the ability of the Bayesian and Winnow classifiers to correct real-word errors will be evaluated.

4.2.1 Evaluation Settings

The two methods are evaluated on confusion sets derived from corrections made to real-word errors during the proofreading phase of Alþingi's digitization project. While there is no theoretical limit to the number of words which may belong to a confusion set, this evaluation is limited to two words per set. This is done both for simplification purposes, and due to the fact that the vast majority of real-word errors occur only between a pair of words. Thus, if a comparison between an unproofread speech and its proofread counterpart reveals that *víð* 'wide' was corrected to *við* 'we' or vice-versa, the confusion set {*við*, *víð*} is created and that particular error is counted among one of its occurrences. In this fashion, a total of 4,344 confusion sets are derived from all 62,555 occurrences of corrected real-word errors.

When the number of errors behind each confusion set is examined, it becomes immediately clear that a vast majority of real-word errors is covered by a relatively small number of confusion sets. For example, {*við*, *víð*} by itself covers 4,979 (7.96%) of all the real-word errors which were corrected. The ratio of real-word errors which are covered by the confusion sets is shown in figure 4.1.

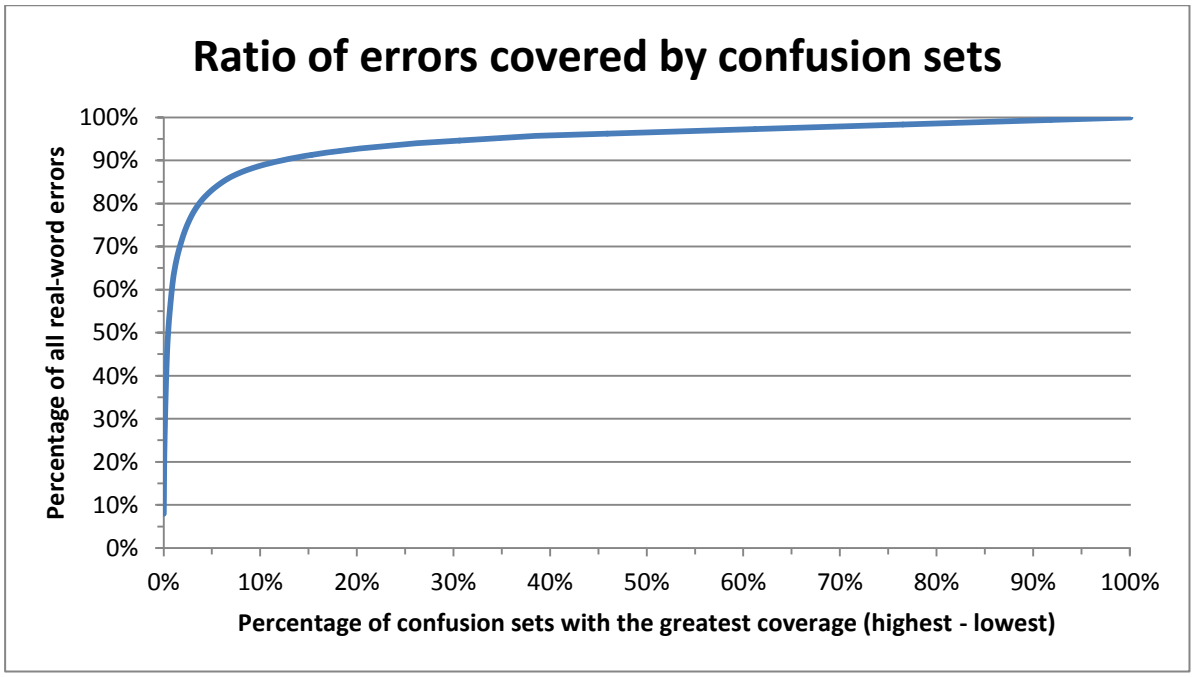

Figure 4.1 The ratio of real-word errors covered by confusion sets

The analysis shows that roughly 80% (50,048) of the real-word errors are covered by only 3.68% (160) of the confusion sets. It further shows that 10% (434) of the confusion sets cover 88.78% (55,538) of all real-word errors. For real-world applications, it may not be feasible to train classifiers for each of the 4,344 confusion sets. The large number of features which the classifiers are trained on may mean that the storage requirements of the training data and the memory requirements of the spellchecker are too high, especially if the spellchecker is to be installed and run on a desktop machine (as opposed to a server). However, as is shown in figure 4.1, it is still possible to cover most of the real-

word errors by focusing only on a relatively small number of confusion sets. For this reason, only errors in the 160 most common confusion sets will be corrected in this evaluation.

Table 4.3 The 25 confusion sets with the greatest real-word error coverage

w_1	w_2	Cov.	C_{Total}	C_{w_1}	C_{w_2}	F_{w_1}	F_{w_2}	E_{w_1}	E_{w_2}
við	víð	7.96%	4979	7	4972	559644	5574	160923	76
sinum	sínum	6.17%	3861	3861	0	4465	24663	13	11286
siðan	síðan	6.03%	3770	3770	0	4016	28019	1	12082
sina	sína	3.09%	1933	1929	4	2434	12694	6	7912
sinu	sínu	2.66%	1665	1665	0	1915	10331	35	5246
mái	mál	2.41%	1509	1508	1	1754	92193	0	4059
sin	sín	2.29%	1431	1431	0	1849	9649	17	4756
siðar	síðar	2.17%	1358	1357	1	1638	9854	8	5876
viða	víða	1.91%	1193	1186	7	1588	4717	9	2718
litið	lítið	1.90%	1187	996	191	6308	9324	1892	3872
siður	síður	1.86%	1163	1162	1	1811	11432	182	2846
visað	vísað	1.68%	1048	1048	0	1095	7343	0	734
sinar	sínar	1.49%	930	926	4	1144	6081	16	2346
tei	tel	1.44%	898	896	2	1005	41727	20	774
bað	það	1.41%	880	878	2	1989	1042117	1346	129673
vei	vel	1.35%	846	845	1	1041	31872	6	15509
lita	líta	1.11%	697	695	2	931	5595	59	2141
vili	vill	0.97%	609	609	0	696	9908	1	3951
álit	álít	0.88%	550	507	43	5552	3293	260	19
visa	vísa	0.81%	505	505	0	525	3920	26	864
liður	líður	0.80%	502	372	130	4117	2555	381	1170
biða	bíða	0.77%	484	482	2	697	3954	1	1251
vist	víst	0.75%	468	462	6	726	3905	174	1676
öli	öll	0.73%	454	453	1	811	17921	14	6040
ríma	tíma	0.63%	391	391	0	412	41470	32	11680

Table 4.3 shows the 25 confusion sets with the greatest coverage of real-word errors. In total, they cover 53.25% of all the real-word errors which were corrected during the proofreading phase. Here, C_{Total} refers to the total number of times that w_1 was changed to w_2 by proofreaders or vice-versa, while C_{w_i} is the number of times that w_i was changed to the other word. F_{w_i} is the frequency with which w_i appears in the unproofread documents. E_{w_i} is the number of sentences in the MÍM corpus containing w_i which could be extracted for use as training examples.

The table shows that corrections within a confusion set tend to go one way but not the other. There are only three confusion sets where more than 10 corrections were made to both words. Also, there is a severe lack of training examples available for word forms in many of the confusion sets above. Without enough training examples for a particular

word form, there is little hope of a classifier being able to learn what context it tends to appear in.

One of the word forms for which training examples are very scarce is *biða* ‘vat’, which is almost never used in modern Icelandic except in the phrase *í belg og biðu* ‘in a torrent’. Inflectional forms other than *biðu* are rarely seen and might not necessarily be recognized by native speakers as being well-spelled word forms. Indeed, the only occurrence of *biða* in the MÍM corpus is from an article describing the meaning and origin of *biðu* from the above phrase. Yet these are all inflectional forms of a word which sees some usage in modern Icelandic, and thus they have their place in the DMII. This means that they are also a part of the lexicon used by the spellchecker, which is problematic because the chance of *biða* legitimately appearing in the Alþingi documents is miniscule compared to it appearing as a misspelling of another word. In this case, it is so unlikely to appear that it may be preferable to remove the word form from the lexicon rather than to try to expand the training corpus to include more examples for it. After all, *bíða* is more than a thousand times more likely to appear in the MÍM corpus than *biða*. A simple baseline classifier that always guesses that *bíða* is correct would have almost perfect accuracy. Even if the Bayesian and Winnow classifiers could attain the same degree of accuracy, they are still much more performance intensive than the simple baseline classifier and may require some additional effort to be undertaken to add more training examples. The same applies to most of the other word forms with a very low number of training examples available. Due to this reason, the Winnow and Bayesian classifiers were only trained on confusion sets where there are at least 30 training examples available for each confusion word. Words with fewer than 30 training examples were classified by the baseline classifier.

Sentences containing confusion words are extracted from portions of the MÍM corpus which can be assumed to be mostly free of spelling errors. This includes text from books and newspaper articles, as well as certain online articles but excludes text from blog posts. In total, this amounts to approximately 15.5 million running words. These sentences serve as training examples for the Winnow and Bayesian classifiers. The methods are then evaluated using 10-fold stratified cross-validation. A maximum of 5,000 examples are used for each confusion word.

4.2.2 Correction of Real-word Errors

The results of the evaluation on the accuracy of the Winnow and Bayesian classifiers are shown in table 4.4. The evaluation extends only to the 160 confusion sets with the greatest coverage where each word has 30 or more training examples available. A further 5 confusion sets, {*af*, *at*}, {*an*, *án*}, {*er*, *et*}, {*og*, *or*} and {*sean*, *sem*}, are excluded, as each contains a very uncommon word in Icelandic (underlined) where the majority of training examples are actually for identically spelled foreign words. In total, 48 confusion sets, covering 26.41% of all real-word errors, are evaluated.

Table 4.4 The classification accuracy of Winnow and Bayesian classifiers

w_1	w_2	Coverage	E_1	E_2	Baseline	Bayesian	Winnow
við	víð	7.96%	5000	76	98.50%	99.47%	99.43%
sinu	sínu	2.66%	35	5000	99.30%	99.72%	99.74%
litið	lítið	1.90%	1892	3872	67.18%	95.02%	96.32%
siður	síður	1.86%	182	2846	93.99%	98.71%	98.81%
bað	það	1.41%	1346	5000	78.79%	97.48%	97.95%
lita	líta	1.11%	59	2141	97.32%	97.40%	97.44%
liður	líður	0.80%	381	1170	75.44%	98.84%	98.84%
vist	víst	0.75%	174	1676	90.59%	97.57%	97.95%
ríma	tíma	0.63%	32	5000	99.36%	99.56%	99.54%
viðar	víðar	0.59%	65	553	89.48%	95.15%	96.28%
mál	mát	0.55%	4030	50	98.77%	99.51%	99.56%
liða	líða	0.52%	271	708	72.32%	97.85%	97.34%
nái	nál	0.43%	482	53	90.09%	96.07%	95.33%
ætla	ætta	0.40%	2675	31	98.85%	99.15%	99.52%
lit	lít	0.39%	666	221	75.08%	98.31%	98.20%
litur	lítur	0.28%	164	873	84.19%	97.11%	97.01%
lifi	lífi	0.25%	162	3086	95.01%	98.58%	98.65%
litum	lítum	0.24%	348	233	59.90%	91.57%	91.74%
hafa	hafi	0.21%	5000	5000	50.00%	81.39%	83.03%
bann	hann	0.19%	542	5000	90.22%	97.80%	97.73%
ber	her	0.18%	3286	219	93.75%	97.92%	97.97%
lagi	tagi	0.18%	3288	1400	70.14%	98.95%	99.47%
há	þá	0.17%	321	5000	93.97%	97.52%	97.26%
vik	vík	0.17%	85	145	63.04%	94.35%	94.35%
bar	þar	0.16%	2057	5000	70.85%	95.39%	95.99%
vikur	víkur	0.16%	984	153	86.54%	97.10%	97.80%
stiga	stíga	0.15%	336	390	53.72%	94.90%	96.28%
vil	vit	0.15%	1470	333	81.53%	99.00%	99.22%
litinn	lítinn	0.14%	169	499	74.70%	96.56%	97.31%
álita	álíta	0.14%	68	90	56.96%	98.10%	100.00%
á	í	0.13%	5000	5000	50.00%	76.75%	76.08%
risa	rísa	0.12%	32	287	89.97%	95.61%	96.24%
að	eð	0.11%	5000	269	94.89%	99.26%	99.66%
liði	líði	0.11%	811	170	82.67%	87.56%	90.11%
sitt	sítt	0.11%	5000	60	98.81%	99.66%	99.57%
lina	lína	0.10%	57	159	73.61%	88.89%	88.89%
sviði	svíði	0.10%	1978	45	97.78%	99.80%	99.80%
benda	henda	0.09%	1665	238	87.49%	95.80%	96.27%
bil	bíl	0.09%	993	818	54.83%	97.63%	97.74%
byggja	hyggja	0.09%	1517	159	90.51%	96.42%	96.78%
bent	hent	0.08%	1287	172	88.21%	94.72%	96.92%
best	hest	0.08%	2646	140	94.97%	97.85%	97.95%
deila	della	0.08%	490	30	94.23%	97.31%	96.35%

w_1	w_2	Coverage	E_1	E_2	Baseline	Bayesian	Winnow
mála	máta	0.08%	1292	176	88.01%	95.16%	95.37%
rann	raun	0.08%	602	3171	84.04%	98.81%	99.52%
sein	sem	0.08%	45	5000	99.11%	99.64%	99.56%
ætlar	ættar	0.08%	996	83	92.31%	98.33%	98.61%
bær	þær	0.07%	128	5000	97.50%	98.87%	98.91%
Average					**83.51%**	**96.34%**	**96.67%**

Here, w_i is a word in a confusion set and E_i is the number of examples available for w_i. Baseline is a classifier which always chooses the word which has the most examples.

For the remaining 112 confusion sets (which cover 53.60% of all real-word errors), the baseline classifier has an average classification accuracy of 98.73%. This affirms that it is preferable to remove these very infrequent words from the lexicon, as such a small portion of them is actually correct (as opposed to the more common words classified by the Winnow and Bayesian classifiers).

Together, the Winnow and baseline classifiers cover 50,048 (80.01%) of the real-word errors in the parliamentary speeches. Assuming that Winnow's average classification accuracy holds, it could be expected to correct 15,971 real-word errors. However, the accuracy of the Winnow classifier may drop when tested on a different corpus from the one it was trained on. Golding and Roth (1999) found that the average classification accuracy dropped from 96.4% to 95.2% when testing on a different corpus. This may mean that assuming an average accuracy of 96.67% may be optimistic.

Additionally, the baseline classifier is able to correct 33,069 real-word errors. Using the Winnow and baseline classifiers in conjunction with nonword error correction on the Alþingi digitization project yields the following expected results:

Table 4.5 Results of nonword and real-word error correction

Error Category	Total errors	Detected	Corrected
Nonword errors	959,052	99.72%	92.90%
Real-word errors	62,555	78.40%	78.40%
Total	**1,021,607**	**98.41%**	**92.01%**

When compared to using nonword correction only, the error detection ratio rises from 93.62% to 98.41%. Furthermore, the ratio of successfully corrected errors rises from 87.21% to 92.01%.

5 Conclusion

The purpose of this thesis was to identify methods for spelling correction that could be modified so as to make them viable for use on Icelandic OCR text and to evaluate them on parliamentary speeches from the Alþingi digitization project.

In the context of spelling correction, the greatest challenge posed by Icelandic is one of data scarcity. It is brought on in large part by the rich morphology of the language, which tends to increase the amount of data required for data-driven NLP tasks, and its small language community, which means that a relatively small amount of language data is being produced.

An important issue caused by data scarcity is the difficulty involved in constructing a lexicon (i.e. a list of well-spelled word forms) with good coverage of the vocabulary. Despite the fact that the lexicon of inflectional forms used in this project contains approximately 2.8 million word forms, more than a third of the out-of-lexicon word forms in the Alþingi digitization project are in fact well-spelled compounds. This gives cause for concern, as relying solely on a lexicon to detect nonword errors means that a great number of false positives (words mistakenly identified as misspellings) are likely to be generated, which is a highly undesirable property for a spellchecker to have. This issue was largely mitigated by the creation and application of a decompounder, which identifies well-spelled compounds among unknown word forms.

With the primary challenges identified, several different methods for spelling correction were reviewed. For nonword error correction, a method based on the one described by Brill and Moore (2000) was proposed. Their method is highly relevant to this work, as it both takes multi-character errors (of which there is a considerable number in the parliamentary speeches) into consideration, and learns the probability of errors occurring from a training set of spelling errors, which can easily be extracted from the digitized parliamentary speeches.

For real-word errors, several methods, categorized as being statistical language model (SLM) based, disambiguation-based and rule-based, were reviewed. While SLM-based methods are very flexible (not requiring real-word errors to be defined beforehand), they are also data-driven and require a substantial amount of language data. Rule-based and disambiguation-based methods were found to be better suited to Icelandic. For this reason, only disambiguation methods were chosen for evaluation, a Winnow classifier (Golding and Roth 1999) and a Bayesian classifier (Golding 1995).

The nonword correction algorithm successfully corrected 93.16% of all errors which it found. This is comparable to results reported by Brill and Moore (2000), who achieved a correction ratio of 93.6%, also without the use of a language model or positional information for character errors. However, their tests were performed on human spelling errors in English as opposed to OCR errors in Icelandic. Additionally, Brill and Moore

trained their model on a list of 10,000 common English misspellings, while in this work the model was trained on 959,052 misrecognized words. It is possible that the considerably larger training set may have mitigated any increased difficulty that comes with correcting Icelandic text as opposed to English. Another interesting difference is that in their work, a lexicon of 200,000 word forms was used while in this work a lexicon of 2.8 million word forms was used.

The Winnow and Bayesian algorithms yielded good results as well. The Bayesian classifier achieved an average classification accuracy of 96.34%, while the Winnow classifier achieved an average classification accuracy of 96.67%. These results are similar to those reported by Golding and Roth (1999), who achieved an average classification accuracy of 93.8% with the Bayesian classifier and 96.6% with the Winnow classifier on English text. However, it may be the case that OCR real-word errors tend to appear in slightly more disjoint contexts than human real-word errors. If this is true, then it may be easier for the disambiguation-based methods to correct OCR errors.

Rule-based methods were not evaluated in this work, although they seem to be viable for the correction of Icelandic real-word errors. A considerable benefit of the rule-based method described by Mangu and Brill (1997) is that despite making use of a very small set of rules, its performance is quite close to that of the Winnow classifier. In addition, the rules it generates are human-readable and therefore easy to edit. For future work, it would be interesting to compare the classification accuracy of rule-based methods to disambiguation-based methods. Additionally, they may prove to be a better choice than the simple baseline classifier for confusion sets with small number of training examples, due to the ease of manually editing and adding new rules.

Additionally, while SLM-based methods were determined to be unlikely to yield satisfactory results, it may be worth considering a hybrid method which relies only in part on language models.

The status of Icelandic as a less-resourced language (Rögnvaldsson et al. 2012) did not prevent good results from being achieved. In fact, an abundance of language resources relevant to the task of spelling correction proved to be available. Furthermore, although data sparseness and other challenges mean certain methods may not be as viable for Icelandic as they are for other languages, the same may not apply to all methods.

Bibliography

Alfonseca, E., Bilac, S., & Pharies, S. (2008). German Decompounding in a Difficult Corpus. In A. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing* (Vol. 4919, pp. 128-139): Springer Berlin / Heidelberg.

Bjarnadóttir, K. (2002). *A Short Description of Icelandic Compounds*. Retrieved from http://www.lexis.hi.is/kristinb/comp-short.pdf.

Bjarnadóttir, K. (2005). *Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum.* Reykjavík: Orðabók Háskólans.

Bjarnadóttir, K. (2012a). The Database of Modern Icelandic Inflection. In *Proceedings of Language Technology for Normalization of Less-Resourced Languages, workshop at the 8th International Conference on Language Resources and Evaluation, LREC 2012,* Istanbul, Turkey. [Forthcoming].

Bjarnadóttir, K. (2012b). [Annotated list of Icelandic compounds]. Unpublished raw data.

Braschler, M., Göhring, A., & Schäuble, P. (2003). Eurospider at CLEF 2002. In C. Peters, M. Braschler & J. Gonzalo (Eds.) *Advances in Cross-Language Information Retrieval* (Vol. 2785, pp. 164-174): Springer Berlin / Heidelberg.

Brill, E., & Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,* Hong Kong.

Brown, R. D. (2002). Corpus-Driven Splitting of Compound Words. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translations (TMI-2002).*

Church, K. W., & Gale, W. A. (1991). Probability scoring for spelling correction. *Statistics and Computing, 1*(2), 93-103.

Creed, A., Dennis, I., & Newstead, S. (1988). Effects of display format on proof-reading with VDUs. *Behaviour & Information Technology, 7*(4), 467-478.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171-176.

Fraser, N. (2012). *Diff, Match and Patch library*. http://code.google.com/p/google-diff-match-patch/

Golding, A. R. (1995). A Bayesian hybrid method for context-sensitive spelling correction. In *Proceedings of the Third Workshop on Very Large Corpora,* Boston, MA.

Golding, A. R., & Roth, D. (1999). A Winnow-Based Approach to Context-Sensitive Spelling Correction. *Machine Learning, 34*(1-3), 107-130.

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., & Richter, M. (2007). Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In *Proceedings of NODALIDA-07,* Tartu, Estonia.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., & Loftsson, H. (2012). *The Tagged Icelandic Corpus (MÍM)*. In *Proceedings of Language Technology for Normalization of Less-Resourced Languages, workshop at the 8th International Conference on Language Resources and Evaluation, LREC 2012,* Istanbul, Turkey. [Forthcoming]

Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine, 15*(3/4).

Howard, J. (2012). *Google Begins to Scale Back Its Scanning of Books From University Libraries.* Retrieved May 9th, 2012, from http://chronicle.com/article/Google-Begins-to-Scale-Back/131109/

Ingason, A., Helgadóttir, S., Loftsson, H., & Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using Hierarchy of Linguistic Identities (HOLI). In B. Nordström & A. Ranta (Eds.) *Advances in Natural Language Processing, 6th International Conference on NLP, GoTAL 2008:* Springer, Berlin, 205-216.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*: Pearson Prentice Hall.

Klijn, E. (2008). The current state of art in newspaper digitisation. A market perspective. *D-Lib Magazine, 14*(1/2).

Koehn, P., & Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1,* Budapest, Hungary.

Kohavi, R., Becker, B., & Sommerfield, D. (1997). Improving simple Bayes. In *Proceedings of the European Conference on Machine Learning.*

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Larson, M., Willett, D., Köhler, J., & Rigoll, G. (2000). Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP),* Beijing, China.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady., 10*(8), 707-710.

Littlestone, N. (1988). Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm. *Machine Learning, 2*(4), 285-318.

Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation, 108*(2), 212-261.

Loftsson, H., & Rögnvaldsson, E. (2007). IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of InterSpeech 2007, Special session: "Speech and language technology for less-resourced languages",* Antwerp, Belgium, 1533-1536.

Loftsson, H., Kramarczyk, I., Helgadóttir, S., & Rögnvaldsson, E. (2009). Improving the PoS Tagging Accuracy of Icelandic Text. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. NEALT Proceeding Series 4. Northern European Association for Language Technology (NEALT), Tartu University Library, Tartu, 103-110.

Lovecraft, H. P. (1928). The Call of Cthulhu. *Weird Tales*. http://archive.org/details/CallOfCthulhu

Lovecraft, H. P. (2011). Kall Cthulhu. Þýðing Þorsteinn Mar Gunnlaugsson. http://nordnordursins.is/2011/09/kall-cthulhu-i-islenskri-thydingu/

Mangu, L., & Brill, E. (1997). Automatic Rule Acquisition for Spelling Correction. In D. Fisher (Ed.) *Proceedings of the Fourteenth International Conference on Machine Learning,* Nashville, Tennessee.

Marek, T. (2006). *Analysis of German Compounds Using Weighted Finite State Transducers.* BA Thesis, Universität Tübingen.

Monz, C., & Rijke, M. d. (2001). *Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian*. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.) *Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*.

Myers, E. (1986). An O(ND) Difference Algorithm and Its Variations. *Algorithmica, 1*, 251-266.

Orwant, J. (2010). Google Books: Making All the World's Books Universally Accessible and Useful. *USENIX Annual Technical Conference*, from https://www.usenix.org/conference/usenix-atc-10/google-books-making-all-worlds-books-universally-accessible-and-useful.

Pind, J., Magnússon, F., & Briem, S. (1991). *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.

Pollock, J. J., & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *Communications of the ACM, 27*(4), 358-368.

Rögnvaldsson, E. (1990). *Um orðaröð og færslur í íslensku*. Reykjavík: Málvísindastofnun Háskóla Íslands.

Rögnvaldsson, E., Jóhannsdóttir, K. M., Helgadóttir, S., & Steingrímsson, S. (2012). The Icelandic Language in the Digital Age, from http://www.malfong.is/Malthing/icelandic_lwp.pdf.

Schiller, A. (2005). German Compound Analysis with wfsc. In *Proceedings of the Fifth Internation Workshop of Finite State Methods in Natural Language Processing (FSMNLP),* Helsinki, Finland, 239-246.

Simons, G. F., & Bird, S. (2008). Toward a Global Infrastructure for the Sustainability of Language Resources. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation,* Cebu City, Philippines.

Taycher, L. (2010). Books of the world, stand up and be counted! All 129,864,880 of you. Retrieved May 9th, 2012, from http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html

Veronis, J. (1988). Computerized correction of phonographic errors. *Computers and the Humanities, 22*, 43-56.

White, R. L. (2000). Object Classification as a Data Analysis Tool. In *Astronomical Data Analysis Software & Systems IX, ASP Conference Proceedings*, Vol. 216, Kona, Hawaii, 577-.

Appendix A: Common Nonword Errors

Table A.1 shows a list of the 200 most common nonword errors in the digitized parliamentary speeches. This list includes all nonword errors where a word in the lexicon was transformed into a word which is not. This excludes numbers (e.g., $1984 \rightarrow l984$), and punctuation marks (e.g., $'' \rightarrow$ "). Further excluded are common abbreviations where spaces have been inserted after periods (e.g., $e.t.v. \rightarrow e.\ t.\ v.$). The majority of these errors (which are quite frequent) are typographical errors which have been corrected during the proofreading of the digitized documents. Ratio shows the ratio of a particular nonword error among all nonword errors (of which there is a total of 959,052). Cumulative shows the cumulative ratio of the *n* most common nonword errors.

Table A.1 The 200 most common nonword errors in the digitized Alþingi texts

n	Correct	Nonword error	Count	Ratio	Cumulative
1	í	i	100,722	10.50%	10.50%
2	í	f	40,119	4.18%	14.68%
3	því	þvi	24,567	2.56%	17.24%
4	til	tii	21,082	2.20%	19.44%
5	ríkisstj	rikisstj	9,366	0.98%	20.42%
6	því	þvf	8,422	0.88%	21.30%
7	tíma	tima	5,383	0.56%	21.86%
8	ég	eg	5,033	0.52%	22.38%
9	líka	lika	3,908	0.41%	22.79%
10	ekki	ekkí	3,905	0.41%	23.20%
11	á	a	3,288	0.34%	23.54%
12	ljóst	ijóst	2,999	0.31%	23.85%
13	vil	vii	2,954	0.31%	24.16%
14	til	tll	2,570	0.27%	24.43%
15	fyrir	fyrír	2,122	0.22%	24.65%
16	ríkisins	rikisins	2,094	0.22%	24.87%
17	till	tili	2,072	0.22%	25.09%
18	síðustu	siðustu	2,017	0.21%	25.30%
19	vísu	visu	1,960	0.20%	25.50%
20	ljós	ijós	1,910	0.20%	25.70%
21	slíkt	slikt	1,886	0.20%	25.90%
22	hafi	hafl	1,879	0.20%	26.10%
23	til	tit	1,833	0.19%	26.29%
24	ríkisstj	rfkisstj	1,794	0.19%	26.48%
25	mín	min	1,689	0.18%	26.66%
26	máli	máll	1,624	0.17%	26.83%
27	heldur	heidur	1,529	0.16%	26.99%
28	mínu	minu	1,489	0.16%	27.15%

n	Correct	Nonword error	Count	Ratio	Cumulative
29	mínum	minum	1,480	0.15%	27.30%
30	slík	slik	1,397	0.15%	27.45%
31	síðasta	siðasta	1,362	0.14%	27.59%
32	reykjavík	reykjavik	1,228	0.13%	27.72%
33	ekki	ekkl	1,214	0.13%	27.85%
34	þessi	þessí	1,201	0.13%	27.98%
35	beinlínis	beinlinis	1,131	0.12%	28.10%
36	síst	sist	1,122	0.12%	28.22%
37	verið	veríð	1,085	0.11%	28.33%
38	allt	alit	1,077	0.11%	28.44%
39	vissulega	víssulega	1,057	0.11%	28.55%
40	tíð	tið	991	0.10%	28.65%
41	till	tiil	985	0.10%	28.75%
42	íslands	islands	973	0.10%	28.85%
43	millj	miilj	890	0.09%	28.94%
44	millj	míllj	876	0.09%	29.03%
45	mílna	milna	871	0.09%	29.12%
46	til	fil	856	0.09%	29.21%
47	verði	verðí	850	0.09%	29.30%
48	held	heid	844	0.09%	29.39%
49	millj	milij	839	0.09%	29.48%
50	slíkar	slikar	829	0.09%	29.57%
51	ríkið	rikið	819	0.09%	29.66%
52	væri	værí	805	0.08%	29.74%
53	slíkum	slikum	804	0.08%	29.82%
54	skuli	skull	790	0.08%	29.90%
55	mjög	m jög	763	0.08%	29.98%
56	komið	komíð	758	0.08%	30.06%
57	hl	hi	758	0.08%	30.14%
58	tíma	tfma	753	0.08%	30.22%
59	eins	eíns	744	0.08%	30.30%
60	til	tíl	741	0.08%	30.38%
61	ekki	elcki	735	0.08%	30.46%
62	slíka	slika	716	0.07%	30.53%
63	íslands	fslands	707	0.07%	30.60%
64	þeim	þeím	682	0.07%	30.67%
65	ríkissjóðs	rikissjóðs	675	0.07%	30.74%
66	mína	mina	673	0.07%	30.81%
67	allt	ailt	665	0.07%	30.88%
68	tími	timi	665	0.07%	30.95%
69	ljúka	ijúka	659	0.07%	31.02%
70	þó	þ6	656	0.07%	31.09%
71	hafi	hafí	641	0.07%	31.16%
72	og	ag	639	0.07%	31.23%
73	síðari	siðari	631	0.07%	31.30%

n	Correct	Nonword error	Count	Ratio	Cumulative
74	til	tli	631	0.07%	31.37%
75	víkja	vikja	626	0.07%	31.44%
76	víðs	viðs	626	0.07%	31.51%
77	nauðsynlegt	nauðsyniegt	596	0.06%	31.57%
78	íslensku	islensku	594	0.06%	31.63%
79	slíku	sliku	593	0.06%	31.69%
80	íslenska	islenska	593	0.06%	31.75%
81	þeir	þeír	592	0.06%	31.81%
82	atriði	atríði	588	0.06%	31.87%
83	fyrir	fvrir	587	0.06%	31.93%
84	l	i	579	0.06%	31.99%
85	reykjavíkur	reykjavikur	575	0.06%	32.05%
86	sínum	sfnum	573	0.06%	32.11%
87	síðan	sfðan	570	0.06%	32.17%
88	hefur	hefnr	565	0.06%	32.23%
89	miklu	míklu	564	0.06%	32.29%
90	að	aó	553	0.06%	32.35%
91	stíl	stil	550	0.06%	32.41%
92	jafnvel	jafnvei	542	0.06%	32.47%
93	íslandi	islandi	532	0.06%	32.53%
94	minni	mínni	525	0.05%	32.58%
95	sambandi	samhandi	524	0.05%	32.63%
96	sinni	sínni	519	0.05%	32.68%
97	máli	máh	518	0.05%	32.73%
98	ríkisstj	ríkísstj	518	0.05%	32.78%
99	mikið	míkið	514	0.05%	32.83%
100	framtíðinni	framtiðinni	511	0.05%	32.88%
101	það	hað	508	0.05%	32.93%
102	ríkissjóður	rikissjóður	506	0.05%	32.98%
103	eftir	eftír	501	0.05%	33.03%
104	ekki	eklci	496	0.05%	33.08%
105	miklu	mikiu	491	0.05%	33.13%
106	ríkissjóði	rikissjóði	489	0.05%	33.18%
107	flm	fim	485	0.05%	33.23%
108	hefur	befur	484	0.05%	33.28%
109	forseti	forsetl	477	0.05%	33.33%
110	eftir	eftlr	473	0.05%	33.38%
111	mílur	milur	469	0.05%	33.43%
112	þannig	þannígg	467	0.05%	33.48%
113	eðlilegt	eðlllegt	465	0.05%	33.53%
114	þrír	þrir	460	0.05%	33.58%
115	dálítið	dálitið	452	0.05%	33.63%
116	til	ti1	452	0.05%	33.68%
117	málið	mállð	446	0.05%	33.73%
118	fyrir	fyrlr	445	0.05%	33.78%

n	Correct	Nonword error	Count	Ratio	Cumulative
119	hefði	hefðí	445	0.05%	33.83%
120	vildi	viidi	444	0.05%	33.88%
121	íslenskum	islenskum	444	0.05%	33.93%
122	því	bví	443	0.05%	33.98%
123	yfir	yflr	442	0.05%	34.03%
124	ísl	isl	441	0.05%	34.08%
125	þess	bess	437	0.05%	34.13%
126	ári	árí	435	0.05%	34.18%
127	við	vlð	434	0.05%	34.23%
128	tímabili	timabili	434	0.05%	34.28%
129	hjá	h já	433	0.05%	34.33%
130	íslandi	fslandi	431	0.04%	34.37%
131	millj	mlllj	431	0.04%	34.41%
132	forseti	forsetí	430	0.04%	34.45%
133	fyrrv	fym	427	0.04%	34.49%
134	máli	málí	420	0.04%	34.53%
135	milli	milll	416	0.04%	34.57%
136	ljósi	ijósi	412	0.04%	34.61%
137	líka	lfka	410	0.04%	34.65%
138	hér	bér	403	0.04%	34.69%
139	slíkri	slikri	402	0.04%	34.73%
140	verið	verlð	399	0.04%	34.77%
141	íslendingar	islendingar	399	0.04%	34.81%
142	apríl	april	398	0.04%	34.85%
143	hafa	bafa	393	0.04%	34.89%
144	íslenskra	islenskra	392	0.04%	34.93%
145	júní	júni	389	0.04%	34.97%
146	framsfl	framsfi	388	0.04%	35.01%
147	um	nm	383	0.04%	35.05%
148	ríkisins	rfkisins	381	0.04%	35.09%
149	telja	teija	381	0.04%	35.13%
150	alþingi	afþingi	378	0.04%	35.17%
151	sérstaklega	sérstakiega	378	0.04%	35.21%
152	þá	bá	375	0.04%	35.25%
153	liggur	llggur	375	0.04%	35.29%
154	sambandi	sambandí	374	0.04%	35.33%
155	öllum	öilum	370	0.04%	35.37%
156	alls	alis	370	0.04%	35.41%
157	allra	alira	370	0.04%	35.45%
158	lögum	iögum	370	0.04%	35.49%
159	ég	l;g	369	0.04%	35.53%
160	heldur	hetdur	366	0.04%	35.57%
161	framtíð	framtið	362	0.04%	35.61%
162	þessari	þessarí	361	0.04%	35.65%
163	fleiri	fieiri	360	0.04%	35.69%

n	Correct	Nonword error	Count	Ratio	Cumulative
164	maí	mai	359	0.04%	35.73%
165	milli	mílli	359	0.04%	35.77%
166	sjálfstfl	sjálfstfi	359	0.04%	35.81%
167	íslendinga	islendinga	355	0.04%	35.85%
168	tímann	timann	353	0.04%	35.89%
169	til	tií	353	0.04%	35.93%
170	milli	milh	350	0.04%	35.97%
171	líkur	likur	349	0.04%	36.01%
172	sinni	sinní	348	0.04%	36.05%
173	litlu	lítlu	348	0.04%	36.09%
174	íslendingar	fslendingar	347	0.04%	36.13%
175	auðvitað	auðvítað	345	0.04%	36.17%
176	líklega	liklega	344	0.04%	36.21%
177	brbl	brbi	339	0.04%	36.25%
178	öðruvísi	öðruvisi	339	0.04%	36.29%
179	hins	híns	338	0.04%	36.33%
180	það	pað	336	0.04%	36.37%
181	alþfl	alþfi	336	0.04%	36.41%
182	gífurlega	gifurlega	336	0.04%	36.45%
183	olíu	oliu	335	0.03%	36.48%
184	tímum	timum	334	0.03%	36.51%
185	sína	sfna	332	0.03%	36.54%
186	sem	aem	328	0.03%	36.57%
187	talið	tallð	327	0.03%	36.60%
188	tíðkast	tiðkast	326	0.03%	36.63%
189	þeirra	þeírra	326	0.03%	36.66%
190	íslendinga	fslendinga	323	0.03%	36.69%
191	yrði	yrðí	321	0.03%	36.72%
192	alþingi	aiþingi	320	0.03%	36.75%
193	en	én	319	0.03%	36.78%
194	einnig	einníg	318	0.03%	36.81%
195	sagði	sagðí	315	0.03%	36.84%
196	skal	skai	314	0.03%	36.87%
197	í	t	313	0.03%	36.90%
198	að	sð	313	0.03%	36.93%
199	efni	efní	313	0.03%	36.96%
200	landinu	iandinu	312	0.03%	36.99%

Appendix B: Common Real-word Errors

Table B.1 shows a list of the 200 most common real-word errors in the digitized parliamentary speeches. This list includes all words in the digitized texts which are in the lexicon and were changed by proofreaders into another word which is also in the lexicon. Ratio shows the ratio of a particular real-word error among all real-word errors (of which there is a total of 62,555). Cumulative shows the cumulative ratio of the *n* most common real-word errors.

Table B.1 The 200 most common real-word errors in the digitized Alþingi texts

n	Word 1	Word 2	Count	Ratio	Cumulative
1	við	víð	4,979	7.96%	7.96%
2	sinum	sínum	3,861	6.17%	14.13%
3	siðan	síðan	3,770	6.03%	20.16%
4	sina	sína	1,933	3.09%	23.25%
5	sinu	sínu	1,665	2.66%	25.91%
6	mái	mál	1,509	2.41%	28.32%
7	sin	sín	1,431	2.29%	30.61%
8	siðar	síðar	1,358	2.17%	32.78%
9	viða	víða	1,193	1.91%	34.69%
10	litið	lítið	1,187	1.90%	36.59%
11	siður	síður	1,163	1.86%	38.44%
12	visað	vísað	1,048	1.68%	40.12%
13	sinar	sínar	930	1.49%	41.61%
14	tei	tel	898	1.44%	43.04%
15	bað	það	880	1.41%	44.45%
16	vei	vel	846	1.35%	45.80%
17	lita	líta	697	1.11%	46.92%
18	vili	vill	609	0.97%	47.89%
19	álit	álít	550	0.88%	48.77%
20	visa	vísa	505	0.81%	49.58%
21	liður	líður	502	0.80%	50.38%
22	biða	bíða	484	0.77%	51.15%
23	vist	víst	468	0.75%	51.90%
24	öli	öll	454	0.73%	52.63%
25	ríma	tíma	391	0.63%	53.25%
26	að	áð	376	0.60%	53.85%
27	viss	víss	375	0.60%	54.45%
28	sig	síg	372	0.59%	55.05%
29	viðar	víðar	371	0.59%	55.64%
30	vissu	víssu	357	0.57%	56.21%
31	vissum	víssum	351	0.56%	56.77%

n	Word 1	Word 2	Count	Ratio	Cumulative
32	mál	mát	347	0.55%	57.33%
33	siðast	síðast	336	0.54%	57.86%
34	liða	líða	327	0.52%	58.39%
35	slikra	slíkra	326	0.52%	58.91%
36	arið	árið	322	0.51%	59.42%
37	betta	þetta	321	0.51%	59.93%
38	vil	víl	294	0.47%	60.40%
39	gislason	gíslason	283	0.45%	60.86%
40	júli	júlí	273	0.44%	61.29%
41	nái	nál	270	0.43%	61.72%
42	viðast	víðast	253	0.40%	62.13%
43	ætla	ætta	251	0.40%	62.53%
44	lit	lít	243	0.39%	62.92%
45	ern	eru	224	0.36%	63.28%
46	bessu	þessu	214	0.34%	63.62%
47	lið	líð	195	0.31%	63.93%
48	linu	línu	181	0.29%	64.22%
49	feist	felst	180	0.29%	64.51%
50	eins	ems	179	0.29%	64.79%
51	það	þáð	178	0.28%	65.08%
52	litur	lítur	177	0.28%	65.36%
53	gripa	grípa	174	0.28%	65.64%
54	mig	míg	164	0.26%	65.90%
55	máli	máti	161	0.26%	66.16%
56	lifi	lífi	159	0.25%	66.41%
57	bæði	hæði	155	0.25%	66.66%
58	lif	líf	151	0.24%	66.90%
59	litum	lítum	149	0.24%	67.14%
60	heist	helst	141	0.23%	67.37%
61	beim	þeim	140	0.22%	67.59%
62	vissi	víssi	139	0.22%	67.81%
63	litt	lítt	133	0.21%	68.02%
64	hess	þess	132	0.21%	68.24%
65	hafa	hafi	131	0.21%	68.45%
66	bessum	þessum	129	0.21%	68.65%
67	feila	fella	124	0.20%	68.85%
68	bessi	þessi	123	0.20%	69.05%
69	vissa	víssa	119	0.19%	69.24%
70	alli	allt	119	0.19%	69.43%
71	bann	hann	116	0.19%	69.61%
72	hetta	þetta	115	0.18%	69.80%
73	áliti	álíti	114	0.18%	69.98%
74	ber	her	112	0.18%	70.16%
75	grein	grem	112	0.18%	70.34%
76	vikið	víkið	111	0.18%	70.51%

n	Word 1	Word 2	Count	Ratio	Cumulative
77	lagi	tagi	110	0.18%	70.69%
78	sén	séu	108	0.17%	70.86%
79	vik	vík	107	0.17%	71.03%
80	málum	mátum	106	0.17%	71.20%
81	bjá	hjá	105	0.17%	71.37%
82	há	þá	104	0.17%	71.54%
83	bessa	þessa	104	0.17%	71.70%
84	bera	hera	103	0.16%	71.87%
85	vita	víta	103	0.16%	72.03%
86	þan	þau	103	0.16%	72.20%
87	bar	þar	100	0.16%	72.36%
88	vissar	víssar	99	0.16%	72.52%
89	vikur	víkur	99	0.16%	72.67%
90	vil	vit	95	0.15%	72.83%
91	breyta	hreyta	94	0.15%	72.98%
92	stiga	stíga	93	0.15%	73.12%
93	vissan	víssan	93	0.15%	73.27%
94	litinn	lítinn	89	0.14%	73.42%
95	álita	álíta	87	0.14%	73.55%
96	láta	táta	86	0.14%	73.69%
97	ætið	ætíð	85	0.14%	73.83%
98	liðið	líðið	85	0.14%	73.96%
99	bara	hara	84	0.13%	74.10%
100	sima	síma	82	0.13%	74.23%
101	álft	álít	82	0.13%	74.36%
102	litra	lítra	81	0.13%	74.49%
103	máls	máts	81	0.13%	74.62%
104	á	í	80	0.13%	74.75%
105	breytt	hreytt	80	0.13%	74.87%
106	málið	mátið	78	0.12%	75.00%
107	petta	þetta	77	0.12%	75.12%
108	hið	híð	75	0.12%	75.24%
109	risa	rísa	74	0.12%	75.36%
110	leið	teið	72	0.12%	75.48%
111	vikum	víkum	72	0.12%	75.59%
112	bila	bíla	72	0.12%	75.71%
113	sitt	sítt	71	0.11%	75.82%
114	er	et	71	0.11%	75.93%
115	að	eð	71	0.11%	76.05%
116	liði	líði	70	0.11%	76.16%
117	linur	línur	69	0.11%	76.27%
118	vorn	voru	69	0.11%	76.38%
119	breyting	hreyting	69	0.11%	76.49%
120	málsins	mátsins	69	0.11%	76.60%
121	lifið	lífið	65	0.10%	76.70%

n	Word 1	Word 2	Count	Ratio	Cumulative
122	davið	davíð	65	0.10%	76.81%
123	an	án	64	0.10%	76.91%
124	lina	lína	64	0.10%	77.01%
125	sviði	svíði	62	0.10%	77.11%
126	milli	milti	62	0.10%	77.21%
127	liðum	líðum	62	0.10%	77.31%
128	smiði	smíði	62	0.10%	77.41%
129	slita	slíta	61	0.10%	77.51%
130	lifa	lífa	59	0.09%	77.60%
131	byggja	hyggja	59	0.09%	77.69%
132	bessar	þessar	58	0.09%	77.79%
133	ætlað	ættað	58	0.09%	77.88%
134	benda	henda	58	0.09%	77.97%
135	vissir	víssir	56	0.09%	78.06%
136	bil	bíl	54	0.09%	78.15%
137	hniga	hníga	53	0.08%	78.23%
138	skina	skína	53	0.08%	78.32%
139	börn	hörn	52	0.08%	78.40%
140	feilur	fellur	52	0.08%	78.48%
141	lifað	lífað	52	0.08%	78.57%
142	deila	della	52	0.08%	78.65%
143	af	at	51	0.08%	78.73%
144	sein	sem	51	0.08%	78.81%
145	breyt	hreyt	50	0.08%	78.89%
146	ætlar	ættar	49	0.08%	78.97%
147	rann	raun	49	0.08%	79.05%
148	rikti	ríkti	49	0.08%	79.13%
149	best	hest	48	0.08%	79.21%
150	mála	máta	48	0.08%	79.28%
151	rannar	raunar	47	0.08%	79.36%
152	bent	hent	47	0.08%	79.43%
153	eiga	elga	46	0.07%	79.51%
154	og	or	45	0.07%	79.58%
155	bær	þær	45	0.07%	79.65%
156	vafi	vafl	45	0.07%	79.72%
157	bilum	bílum	45	0.07%	79.79%
158	sean	sem	45	0.07%	79.87%
159	kama	koma	44	0.07%	79.94%
160	eina	ema	44	0.07%	80.01%
161	akkar	okkar	44	0.07%	80.08%
162	ein	em	43	0.07%	80.15%
163	par	þar	43	0.07%	80.21%
164	tif	til	42	0.07%	80.28%
165	borið	horið	42	0.07%	80.35%
166	kari	karl	41	0.07%	80.41%

n	Word 1	Word 2	Count	Ratio	Cumulative
167	hó	þó	40	0.06%	80.48%
168	rími	tími	40	0.06%	80.54%
169	frá	írá	40	0.06%	80.61%
170	vitt	vítt	39	0.06%	80.67%
171	beri	heri	39	0.06%	80.73%
172	liðir	líðir	39	0.06%	80.79%
173	kviða	kvíða	39	0.06%	80.86%
174	biður	bíður	38	0.06%	80.92%
175	öllum	ötlum	38	0.06%	80.98%
176	smiða	smíða	37	0.06%	81.04%
177	visi	vísi	37	0.06%	81.10%
178	liti	líti	37	0.06%	81.15%
179	mal	mál	37	0.06%	81.21%
180	landinn	landinu	36	0.06%	81.27%
181	hví	því	36	0.06%	81.33%
182	ar	er	36	0.06%	81.39%
183	gislasonar	gíslasonar	35	0.06%	81.44%
184	farið	fárið	34	0.05%	81.50%
185	og	op	33	0.05%	81.55%
186	veita	velta	33	0.05%	81.60%
187	veit	velt	33	0.05%	81.65%
188	tina	tína	33	0.05%	81.71%
189	bægt	hægt	33	0.05%	81.76%
190	siðla	síðla	33	0.05%	81.81%
191	bilar	bílar	33	0.05%	81.87%
192	bíða	híða	32	0.05%	81.92%
193	hím	hún	32	0.05%	81.97%
194	byrja	hyrja	32	0.05%	82.02%
195	af	al	32	0.05%	82.07%
196	sviðum	svíðum	32	0.05%	82.12%
197	her	hér	32	0.05%	82.17%
198	vitum	vítum	32	0.05%	82.22%
199	gripið	grípið	32	0.05%	82.27%
200	pappir	pappír	31	0.05%	82.32%