LIU-IDA/KOGVET-A–12/014–SE

Linköping University

Master Thesis

Automatic Text Simplification via Synonym Replacement

by

Robin Keskisärkkä

Supervisor: **Arne Jönsson**
Dept. of Computer and Information Science
at Linköping University

Examinor: **Sture Hägglund**
Dept. of Computer and Information Science
at Linköping University

Abstract

In this study automatic lexical simplification via synonym replacement in Swedish was investigated using three different strategies for choosing alternative synonyms: based on word frequency, based on word length, and based on level of synonymy. These strategies were evaluated in terms of standardized readability metrics for Swedish, average word length, proportion of long words, and in relation to the ratio of errors (type A) and number of replacements. The effect of replacements on different genres of texts was also examined. The results show that replacement based on word frequency and word length can improve readability in terms of established metrics for Swedish texts for all genres but that the risk of introducing errors is high. Attempts were made at identifying criteria thresholds that would decrease the ratio of errors but no general thresholds could be identified. In a final experiment word frequency and level of synonymy were combined using predefined thresholds. When more than one word passed the thresholds word frequency or level of synonymy was prioritized. The strategy was significantly better than word frequency alone when looking at all texts and prioritizing level of synonymy. Both prioritizing frequency and level of synonymy were significantly better for the newspaper texts. The results indicate that synonym replacement on a one-to-one word level is very likely to produce errors. Automatic lexical simplification should therefore not be regarded a trivial task, which is too often the case in research literature. In order to evaluate the true quality of the texts it would be valuable to take into account the specific reader. A simplified text that contains some errors but which fails to appreciate subtle differences in terminology can still be very useful if the original text is too difficult to comprehend to the unassisted reader.

Keywords : Lexical simplification, synonym replacement, SynLex

Acknowledgements

This work would not have been possible without the support of a number of people. I would especially like to thank my supervisor Arne Jönsson for his patience and enthusiasm throughout the entire work. Our discussions about possible approaches to the topic of this thesis have been very inspirational. I would also like to thank Christian Smith for giving me access to his readability metric module, and Maja Schylström for her help as an unbiased rater of the modified texts. A final thanks goes out to Sture Hägglund for his enthusiasm and support in the beginning stages of this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The field of automatic simplification of text has been gaining momentum over the last 20 years. Modern developments in computer power, natural language processing tools, and increased availability of corpora are a few advancements which have made many modern efforts possible.

The motivating factors for text simplification are abundant. For example, in one study 25 percent of the adult Swedish population were shown to have difficulties with reading and comprehending newspaper articles in topics that were unfamiliar to them (Köster-Bergman, 2001), a surprisingly high figure given that the almost the whole population is considered literate (Mühlenbock and Johansson Kokkinakis, 2010). Even text documents that have been created for a specific group of readers can cause problems for people inside the profession (Dana, 2007). To be able to read and properly comprehend complicated texts is of profound importance in countries where instructions and information presented in written form is the norm. The matter is complicated further by the fact that the group in need of specifically adapted information is highly heterogeneous and no single *easy-to-read* text is suitable for all readers (Mühlenbock and Johansson Kokkinakis, 2010). Aaron et al. (1999) showed that poor readers among children could broadly be categorized by deficiencies in decoding, comprehension, a combination of decoding and comprehension, or reading speed and orthographic processing. Though the degree to which this study applies to adults and second language learners is uncertain it can be concluded that the needs of readers vary greatly.

People affected by poor reading skills may not only suffer from aphasia, dyslexia, or cognitive disability, but also include second language learners,

and adults lacking proper schooling. Aside from the literacy skills of the reader motivation, background knowledge, and other factors also affect the ease by which readers decode and comprehend texts (Feng et al., 2009).

A considerable of amount information is unavailable to poor readers since texts may be too difficult, too long, or require a disproportionate amount of effort. Simplified versions of newspaper material, public information, legal documents, and medical resources, to name a few, would enable the majority of these readers to benefit from this information. But manual simplification of documents is very time consuming and therefore very expensive, and despite efforts to make public information more accessible the majority of the texts available do not have any specifically adapted texts for people with reading difficulties.

Attempts have been made to create systems that automatically make texts easier to read. Two common techniques include automatic text summarization systems, which attempt to abstract or extract only the most important sentences or information from a text (Smith and Jönsson, 2007; Luhn, 1958), and syntactic simplification (Siddharthan, 2003; Carroll et al., 1998, 1999). The summarization methods may be seen as text simplification systems since many poor readers have particular problems with long texts. Shorter texts can make information more salient and lessen the amount of effort required to comprehend a text, both for poor and skilled readers. One risk of summarization systems is, however, that they often increase the information density of the text, which can make the text more difficult to read.

Syntactic text simplification techniques involve rewriting texts to create simpler sentence structures. By using part-of-speech tagging rule-based syntactic simplification operations may be applied to individual sentences (Kandula et al., 2010; Chandrasekar et al., 1996; Siddharthan, 2003; Rybing et al., 2010; Decker, 2003). These rewrite rules may, among many other things, split long sentences into shorter ones, rewrite verb form from passive to active, remove superfluous words, or apply anaphora resolution to reduce the readers memory load. Some of these measures have been directly motivated by cognitive factors, while others have been deduced from comparisons of characteristics between texts of varying difficulty.

Other techniques for simplification of text may include adding semantic information to aid the reader (Kandula et al., 2010), replace difficult terminology with simpler synonymous alternatives (Carroll et al., 1999, 1998), and the inclusion of word lists explain central terminology (Kokkinakis et al., 2006).

Manually simplified Swedish text for language impaired readers has received a lot of attention for more than 60 years. For example, *Centrum för lättläst* provides readers with news written in easy read format, and the related publishing company *LL-förlaget* republishes books in easy read formats (http://www.lattlast.se/). However, the vast majority of research in automatic text simplification has been conducted for the English language, and research for Swedish is still scarce in the literature.

1.1 Purpose of the study

The purpose of this study is to investigate automatic lexical simplification in Swedish. Studies within lexical simplification have historically investigated the properties of English mainly, and almost all rely in some way on the use of WordNet (Carroll et al., 1998; Lal and Rüger, 2002; Carroll et al., 1999). WordNet is a resource and research tool that contains a lot of linguistic information about English, such as semantic relations between words and word frequency counts. For Swedish there is no database, tool, or system of similar magnitude or versatility.

A few studies have used lexical simplification as a means of simplifying texts to improve automatic text summarization (Blake et al., 2007), and some have applied some type of lexical simplification coupled with syntactic simplification, but studies that focus on lexical simplification in its own right are rare. The studies that do exist tend to view lexical simplification as a simple task in which words are replaced with simpler synonyms, defining a *simpler* word as one that is more common than the original. Naturally, familiarity and the perceived difficulty of a word is related to how often an individual is exposed to it and thus its frequency, but to the author's knowledge there has been no research concerning what difference in frequency should have to apply for a word to be considered simpler than another. For example, in the Swedish Parole list of frequencies for words *allmän* (general) has a frequency count of 686 and its possible synonym *offentlig* (public) has a frequency count of 604; does this relatively small difference in frequency warrant a replacement? At the same time some words can, despite being quite common, be complicated to read as in the case of *folkomröstning* (referendum), or difficult to comprehend as in the case of *abstrakt* (abstract).

The difficulty of a word in terms of readability is affected by length, often measured in terms of number of syllables, or number of characters. For example, the phonological route may not be used effectively

by individuals with phonological impairment, as is presumed to be the case for some people suffering from dyslexia, and the affects become very prominent for long words. Also, many of the most popular readability metrics use number of letters, or syllables, as a component in estimating the difficulty of texts at the document level.

The aim of the current study is to investigate whether a text can be successfully simplified using synonym replacement on the level of one-to-one word replacement. Theoretically, synonym replacements can affect established readability metrics for Swedish, mainly LIX and OVIX, in different ways. LIX can be affected by changes in the number of long words within the text, and the average word length, while number of words per sentence, and number of sentences remains unchanged. OVIX on the other hand, which is a metric that estimates vocabulary load, can be affected by a change in the variation in vocabulary.

The correlation between word lengths and text difficulty indicates that lexical simplification via replacement is likely to result in decreased word length overall, and a decrease in number of long words, if the text is simplified. Also, if words are replaced by simpler synonyms one could, depending on the technique employed, expect a smaller variation in terms of unique words, since multiple nuanced words may be replaced by the same word. But readability metrics in themselves do not tell the whole story about the actual quality of a text.

There are very few examples of words with identical meaning in all contexts, if any, and any tool that replaces synonyms automatically is likely to accidentally affect the content of the text. This, however, does not unequivocally mean that lexical simplification using synonym replacement would not be useful. For example, individuals with limited knowledge of economy may profit very little by the distinction between the terms *income*, *salary*, *profit*, and *revenue*. Replacing these terms with a single word, say only *income*, would probably result in a document that fails to appreciate the subtle differences between these three concepts, but it does not necessarily affect the individual's understanding of the text to the same degree, especially when the word appears in context.

The aim of the study can be summarized into three main questions:

- To what degree can automatic lexical simplification on the level of one-to-one synonym replacement be successfully applied to Swedish texts?
- How can thresholds for replacements be introduced to maximize the quality of the simplified document?

- What are the major drawbacks of this method, and how can these problems be mitigated?

The study is in many ways exploratory, as the limitations of lexical simplification for Swedish to date are largely unknown. The study will derive some of the central concepts from international research, mainly conducted for the English language, but the resources utilized by the implemented simplification modules rely heavily on the results from existing Swedish research.

Chapter 2

Background

This chapter introduces some of the main concepts and previous research underlying this thesis. Some of the concepts are discussed in some depth, while others are introduced mainly as a way of orienting the reader in the field of automatic text simplification.

2.1 Automatic text simplification

The field of automatic text simplification dates back to the middle of the 1990s. In one early paper Chandrasekar et al. (1996) summarizes some techniques that can be used to simplify the syntax of text, with the primary aim of simplifying complicated sentences for systems relying on natural language input. The simplification processes described would, however, also apply to human readers. They suggest that simplification can be more or less appropriate depending on the context. For example, legal documents contain a lot of nuances of importance, and since simplification may result in a loss some or all of these distinctions this is probably not a suitable context. In other contexts the implications may be less noticeable and be outweighed by the advantages of a simplified document.

Chandrasekar and Srinivas (1997) view simplification as a two-stage process: analysis followed by transformation. Their system works on sentence level simplification is expressed in the form of transformation rules. These rules could be hand-crafted (Decker, 2003) but this process is very time consuming since it has to be repeated for every domain. Using

a set of training data Chandrasekar and Srinivas (1997) automatically induced transformation rules for sentence level simplification.

Carroll et al. (1998) and Carroll et al. (1999) describe the work carried out in a research project called PSET (*Practical Simplification of English Text*). In this project a system was developed explicitly to assist individuals suffering from aphasia in reading English newspaper texts. Although their primary interest was aphasia they suggest that the same system may be generalizable to second language learners as well. The system can be described as a two part system, where the text is first analyzed, using a lexical tagger, a morphological analyzer, and a parser, and then passed to a simplifier. The simplifier consists of two parts: a syntactic simplifier, and a lexical simplifier. This system's architecture is quite similar to that found in (Chandrasekar and Srinivas, 1997).

Kandula et al. (2010) is another study using syntactic transformation rules for simplification of text. This study, however, employed a threshold for sentence length to decide whether a sentence needed simplification. Their threshold was set to ten words, meaning that every sentence longer than ten words was passed through a grammatical simplifier, which could break down sentences into two or more sentences as described by Siddharthan (2003). Apart from using a threshold for simplification of a sentence the study also required every simplified sentence to be at least seven words having noted that shorter sentences often became fragmented and unlikely to improve readability. Two more criteria were used to decide whether a simplified sentence should be accepted: estimation of the *soundness* of sentence's syntax based on link grammar, and the OpenNLP score, where a threshold was established empirically.

Syntactic simplification is not the only means by which people have tried to automatically simplify text. Smith and Jönsson (2007) showed that automatic summarization of Swedish text can increase documents readability. They showed that the summarization affected different genres of texts in slightly different ways, but the results showed an average decrease in LIX-value across all genres for summaries of varying degrees. For some texts there was also a decrease in OVIX, indicating that it is possible for idea density of sentences to decrease when a text is summarized. Since the effort required to read a text generally increases with its length other benefits in terms of readability, not captured by established Swedish readability metrics, also comes from summarizing a text. A third area that can be used as a means of simplifying text is lexical simplification.

2.2 Lexical simplification

Lexical simplification of written text can be accomplished in a variety of ways. Replacement of difficult words and expressions with simpler equivalences is one such strategy. But lexical simplification may also include introduction of explanations or removal of superfluous words.

One way of performing lexical simplification was implemented by Carroll et al. (1998, 1999). Their simplifier used word frequency count to estimate the difficulties of words. Their system passed word one at a time through the WordNet lexical database to find alternatives to the presented word. An estimate of word difficulty was then acquired by querying the Oxford Psycholinguistic Database for the frequency of the word. The word with the highest frequency was selected as the most appropriate word and was used in the reconstructed text. They observed that less frequent words are less likely to be ambiguous than frequent ones since they often have specific meanings.

Lal and Rüger (2002) used a combination of summarization and lexical simplification to simplify a document. Their system was constructed within the GATE framework, which is a modular architecture where components can easily be replaced, combined, and reused. They based their lexical simplification on queries made to WordNet in a fashion very similar to Carroll et al. (1998), and word frequency counts were used as an indicator of word difficulty. No word sense disambiguation was performed, instead the most common sense was used. Their simplification trials were informal and they observed problems both with the sense of the words and with strange sounding language, something they suggest could be alleviated by introducing a collocation look-up table.

Kandula et al. (2010) simplified text by replacing words with low familiarity scores, as identified by a combination of the words usage contexts and its frequency in lay reader targeted biomedical sources. The familiarity score as an estimation of word difficulty was successfully validated using customer surveys. Their definition of familiarity score results in a number within the range of 0 (very hard) and 1 (very easy). The authors employed a threshold of familiarity to decide whether a word needed to be simplified, and alternatives were looked up in a domain specific look-up table for synonyms. Replacements were performed if the alternative word satisfied the familiarity score threshold criterion. If there was no word with sufficiently high familiarity score an explanation was added to the text. The explanation generation based on the relationship between the difficult term and a related term with higher familiarity score would be

used to generate a short explanation phrase. An explanation took either the form <difficulterm> (a type of <parent>) or <difficulterm> (e.g. <child>), depending on the relationship between the two words, but as an earlier study showed these two relations produced useful and correct explanations in 68% of the generated explanations, the authors also introduced non-hierarchical semantic explanation connectors.

Another lexical simplification technique is to remove sections of a sentences that are deemed to be non-essential information, a technique that among other things has been used to simplify text to improve automatic text summarization (Blake et al., 2007).

2.3 Semantic relations between words

The semantic relations between words are often described in terms of synonymy (similar), antonymy (opposite), hyponymy (subordinate), meronymy (part), troponymy (manner), and entailment (Miller, 1995). The last two categories, troponymy and entailment, deal with verb relations specifically. Synonymy and antonymy are frequently used in dictionaries to describe the meaning of words. For example the noun *bike* may described as a synonym to *bicycle* and the preposition *up* may described as the opposite of its antonym *down*. These relationships are not always straightforward and more than one semantic relationship must often be used to specify a word's meaning.

2.3.1 Synonymy

Synonyms can be described as words which have the same or almost the same meaning in some or all senses (Wei et al., 2009), as a symmetric relation between word forms (Miller, 1995), or words that are interchangeable in some class of contexts with insignificant change to the overall meaning of the text (Bolshakov and Gelbukh, 2004). Bolshakov and Gelbukh (2004) also made the distinction between *absolute* and *non-absolute* synonyms. They describe absolute synonyms as words of linguistic equivalence that have the exact same meaning, such as the words in the set {*United States of America*,*United States*, *USA*, *US*}. Absolute synonyms can occur in the same context without significantly affecting the overall style or meaning of the text, but equivalence relations are extremely rare in all languages. Bolshakov and Gelbukh suggested that the inclusion of multiword and compound expressions in synonym databases nevertheless

brings a considerable amount of absolute synonym relations.

A group of words that are considered synonymous are often grouped into synonym sets, or synsets. Each synonym within a synset are considered synonymous with the other words in that particular set (Miller, 1995). This builds on the assumption that that synonymy is a symmetric property, that is, if *car* is synonymous with *vehicle* then *vehicle* should be regarded as synonymous to *car*. Synonymy is commonly also viewed as a transitive property, that is, if $word_1$ is a synonym of $word_2$ and $word_2$ is synonym of $word_3$ then $word_1$ and $word_3$ can be viewed as synonyms (Siddharthan and Copestake, 2002). This view is not entertained in this thesis, since overlapping groups of synonyms can result in extremely large synsets, especially if word sense disambiguation is not applied. The view of synonymy as symmetric and transitive property is seldom discussed in literature but is closely related to the distinction of hyponyms.

Hyponyms express a hierarchical relation between two semantically related words. One example of this is that the synonym pair used in the previous example can be regarded as a hyponym relation, where *car* is a hyponym of *vehicle*, that is, everything that falls within the definition of *car* can also be found within the definition of *vehicle*. Again, just as absolute synonyms are rare so are true hyponym relations, but this distinction raises some questions. These two words can be viewed as synonymous in some cases, but in most cases *vehicle* has a more general meaning than *car*. Replacement of the term *car* for *vehicle* would thus, in most contexts, produce a less precise distinction but would likely not introduce any errors. However, if the opposite were to occur, that is, if *vehicle* would be replaced by *car*, the distinction would become more explicit and would run a higher risk of producing errors. In practise, many words cannot be ordered hierarchically but rather exist on the same level with an overlap of semantic and stylistic meaning.

In WordNet (Miller, 1995) hyponyms are expressed as separate relation from synonyms, and for Swedish a similar hierarchical view of words can be found in the semantic dictionary SALDO (Borin and Forsberg, 2009). SALDO is structured as a lexical-semantic network around two primitive semantic relations. The main descriptor, or *mother*, is closely related to the headword but is more central (often a hyponym or synonym, but sometimes even an antonym). Unlike WordNet SALDO contains both open and closed word classes.

2.4 Readability metrics

To study readability of texts a number of readability metrics have been developed. This section briefly describes the established readability metrics for Swedish and the textual properties that they tend to reflect.

2.4.1 LIX

LIX, läsbarhetsindex (*readability index*), is the most widely used readability metric for Swedish to date. LIX is described by the number of words per sentence and the proportion of long words (>6 characters). Figure 2.1 shows the formula used to calculate the LIX-value of a text.

$$\text{LIX} = \frac{\text{number of words}}{\text{number of sentences}} + \left(\frac{\text{number of words} > 6 \text{ characters}}{\text{number of words}} \times 100 \right)$$

Figure 2.1: The formula used to calculate LIX.

A text's readability given its LIX-value corresponds roughly to a genre as seen in the reference table for readability presented in Table 2.1 (Mühlenbock and Johansson Kokkinakis, 2010).

Table 2.1: Reference readability values for different text genres (Mühlenbock and Johansson Kokkinakis, 2010).

LIX-value	Text genre
–25	Children's books
25–30	Easy texts
30–40	Normal text/fiction
40–50	Informative text
50–60	Specialist literature
>60	Research, dissertations

2.4.2 OVIX

OVIX, ordvariationsindex (word variation index), is a metric that describes vocabulary load by calculating the lexical variation of a text. High

values are typically associated with lower readability. The formula for calculating OVIX is presented in Figure 2.2.

$$\text{OVIX} = \frac{\log(\text{number of words})}{\log\left(2 - \frac{\log(\text{number of unique words})}{\log(\text{number of words})}\right)}$$

Figure 2.2: The formula used to calculate OVIX.

2.4.3 Nominal ratio

Nominal ratio (NR) is calculated by dividing the number of nouns, prepositions, and participles with the number of pronouns, adverbs and verbs. An NR-value of 1.0 is the average level of for example newspaper texts. Higher values reflect more stylistically developed text, while lower values indicate more simple and informal language. Low NR-values can also indicate a more narrative text type (Mühlenbock and Johansson Kokkinakis, 2010). The formula used to calculate NR is presented in Figure 2.3. NR is not affected by synonym replacements since words are replaced in a one-to-one fashion of, presumably, the same word class, and the metric is primarily used in this study as an aid to estimate the readability of texts.

$$\text{NR} = \frac{\text{nouns} + \text{prepositions} + \text{participles}}{\text{pronouns} + \text{adverbs} + \text{verbs}}$$

Figure 2.3: The formula used to calculate nominal ratio (NR).

Chapter 3

A lexical simplification system

This chapter describes the development of a lexical simplification system, which is intended to replace words with simpler synonyms. The chapter describes the implementation of a number of modules, and the motivations of the various techniques that these employ.

3.1 Synonym dictionary

In order to produce a lexical simplification system for synonym replacement one requirement is a list or database containing known synonyms in some form. An interesting resource for synonyms is the freely available SynLex, which is a synonym lexicon containing about 38,000 Swedish synonym pairs. This resource was constructed in a project at KTH by allowing Internet users of the Lexin translation service to rate the strength of possible synonyms on a scale from one to five (Kann and Rosell, 2005). Users were also allowed to suggest their own synonym pairs, but these suggestions were checked manually for spelling errors and obvious attempts at damaging the results before being allowed to enter the research set. The average counts were summarized after a sufficient number of responses had been gathered for each word pair. The list of word pairs was then split into two pieces, retaining all pairs with a *synonymy level* that was equal to, or greater than, three.

3.2 Combining synonyms with word frequency

In order to create a resource of synonym pairs containing synonyms and an account of how frequent each word is in the Swedish language SynLex was combined with Swedish Parole's frequency list of the 100,000 most common words into a single XML-file. This file contained synonym pairs in lemma form, the level of synonymy between the words, and word frequency count for each of the words.

Frequency count was found by taking into consideration the different inflection forms of each word by using the Granska Tagger (Domeij et al., 2000), a part-of-speech tagger for Swedish, to generate the lemma forms of the words in the Parole list. Frequency counts for each identical lemma was then collapsed into a more representative list of word frequencies. The lemma frequencies for words based on this list were then added as an attribute to each word in the synonym XML-file. If the word did not have a frequency count in the Parole-file the entry was excluded from the synonym list.

The original SynLex file (http://folkets2.nada.kth.se/synpairs.xml) contained a total of 37,969 synonym pairs. When adding frequency counts to the lemma forms of these words, and excluding pairs with zero frequency counts for any of the words, 23,836 pairs remained. Fewer synonym pairs may have been lost if the entire Parole frequency count list had been used, rather that limiting it to the 100,000 most common words, but SynLex contained some combination pairs that where not one-to-one word pairings while Parole only has frequency counts for individual words. Another factor which affected the number of synonym pairs was the precision of the Granska Tagger in identifying lemma forms. Table 3.1 shows a portion of the generated synonym XML-file.

Table 3.1: Three examples from the synonym XML-file.

<entry level=”4.0”>		
<word1 freq=”12”>	abdikera	</word1>
<word2 freq=”304”>	avgå	</word2>
<entry level=”3.4”>		
<word1 freq=”2484”>	avgöra	</word1>
<word2 freq=”1381”>	bedöma	</word2>
<entry level=”4.2”>		
<word1 freq=”2484”>	avgöra	</word1>
<word2 freq=”2888”>	besluta	</word2>
</entry>		

3.3 Synonym replacement modules

Three main modules where developed in Java, which given a text input file could generate a new text file in which synonym replacement had been performed. By looking up the possible synonyms for every word in the document the three modules identified the best alternative word based on word frequency, word length, or level of synonymy.

In the first module replacements were motivated by word frequency counts, which have been used to estimate reader familiarity with a word in several studies (see section 2.2). A reader is more likely to be familiar with a word if it is commonly occurring.

Replacements in the second module were motivated by established readability metrics which state that word length correlates with readability of text. By replacing words with shorter alternatives the average word length decreases and, hypothetically, the overall text difficulty of the text decreases. The general idea is that word length is a good estimate of the difficulty of a word.

The third module motivates replacements based on the level of synonymy between the words in SynLex. For all modules the support for threshold criteria was introduced.

3.4 Handling word inflections

The developed modules could originally only replace exact matches to the synonyms in the synonym XML-file. This meant that only words written in their lemma form could be replaced. In order to increase the number of replacements, as well as to handle word class information and word inflections, a simple inflection handler was developed. The Granska Tagger was used to generate a list with inflection patterns for the words in the synonym dictionary. These were stored in a separate specially formatted XML-file. A Java class was developed which, in conjunction with this XML-file, enabled word forms of lemmas to be looked up quickly by passing lemma and inflection notation, which can be generated for a word using the Granska Tagger.

The modules were modified to generate lemma and word class information for each word in the text, and to look for a synonym based on the lemma. If the original class and inflection form could be generated using the inflection handler it was regarded as a possible replacement alternative. Table 3.2 shows a portion of the generated inflection XML-file.

Table 3.2: An example from the word inflection XML-file showing the generated word forms of *mamma* (mother).

```
<word>
<lemma> mamma </lemma>
<alt>
nn.utr.plu.ind.gen=mammors
nn.utr.sin.ind.gen=mammas
nn.utr.sms=mamma
nn.utr.plu.def.nom=mammorna
nn.utr.sin.def.gen=mammans
nn.utr.sin.ind.nom=mamma
nn.utr.plu.ind.nom=mammor
nn.utr.plu.def.gen=mammornas
nn.utr.sin.def.nom=mamman
</alt>
</word>
```

3.5 Open word classes

Synonym replacement is especially prone to errors when not taking into consideration word class information. In order to minimize errors caused by this a filter was appended to the replacement modules which allowed only open word classes to be replaced, i.e. replacements were only performed on words belonging to the word classes nouns, verbs, adjectives, and adverbs. The rationale behind this filter was that the closed words form a group which is only rarely extended and is often related to the structure and form of the sentence, rather than to its specific semantic meaning. Also, the word frequency of the closed word classes is much greater than for words in general, and these words are therefore almost always very familiar to readers, with a few rare exceptions. As an example, in the Swedish Parole corpus the first 30 closed words have a sum of frequency exceeding the collapsed sum of frequencies for all other words down to the 500th most common word.

3.6 Identification of optimal thresholds

For each of the synonym replacement modules described in section 3.3 a threshold for the criteria of a substitution is supported, such that if the criteria value is too low the substitution will not occur. The selection criteria employed by the modules ensures that only the word with the highest criteria value replaces the original word. Introducing a threshold would thus only prune replacements from among the least qualified words in the set being replaced. Raising the threshold sufficiently would eventually stop all substitutions from occurring. Establishing optimal thresholds for the different criteria can therefore be done by a stepwise increase of the threshold criteria. Analyzing the ratio of errors in relation to the number of substitutions could then possibly establish thresholds for the replacements strategies.

Chapter 4

Method

The following chapter describes the methods that were used to evaluate the performance of the modules in the different experiment settings. It also describes and compares the texts which were used in the experiments.

4.1 Selection of texts

In an attempt to cover a variety of different genres texts were selected from four different sources: newspaper articles from *Dagens nyheter*, informative texts from Försakringskassan's homepage, articles from *Forskning och framsteg*, and academic text excerpts. Every genre consisted of four documents of roughly the same size, though the newspaper articles were slightly shorter on average.

4.1.1 Estimating text readability

The established Swedish readability metrics LIX, OVIX, and nominal ratio were used to estimate the difficulty of the four genres (see 2.4 for more information about the readability metrics). The four genres were selected to represent a spectrum of readability, and the documents were hypothesized to represent different readability levels. In terms of readability the texts could, however, not be arranged in any definite order. The academic text excerpts (ACADEMIC), for example, were clearly the most difficult in terms of LIX-value and the articles from *Forskning och framsteg* (FOF) had the highest OVIX-values. The newspaper articles from *Dagens Ny-*

heter (DN) had the lowest LIX-value as well as the lowest nominal ratio among the genres, but had a higher OVIX-value than the informative texts from Försakringskassan's homepage (FOKASS). This inconsistency could possibly be explained by the difference in average text length since OVIX is affected by the length of the text and as a result shorter texts can receive higher OVIX-values.

Table 4.1: Average readability metrics for the genres *Dagens nyheter* (DN), *Försakringskassan* (FOKASS), *Forskning och framsteg* (FOF), *academic text excerpts* (ACADEMIC), and for all texts, with readability metrics LIX (readability index), OVIX (word variation index), and nominal ratio (NR). The table also presents *proportion of long words* (LWP), *average word length* (AWL), *average sentence length* (ASL), and *average number sentences* per text (ANS).

Genre	LIX	OVIX	NR	LWP	AWL	ASL	ANS
ACADEMIC	53	66.5	1.4	0.28	5.1	23.6	51
DN	41	66.4	1.0	0.23	4.7	17.7	43
FOF	44	77.4	1.5	0.27	4.9	16.7	58
FOKASS	44	49.1	1.1	0.26	5.1	17.5	64
All texts	46	64.9	1.3	0.26	5.0	18.9	54

4.2 Analysis of errors

In order to evaluate how often the synonym replacement modules produce erroneous substitutions errors were identified by hand. The distinction of errors can in some cases be subjective, which motivated the use of a predefined manual.

4.2.1 Two types of errors

The techniques employed by the modules can produce a variety of different errors including deviations from the original semantic meaning, replacement of established terminology, formation of strange collocations, deviation from general style, syntactic or grammatical incorrectness and more. For the purpose of this study some of the possible errors were ignored while the remaining were clustered into two separate categories:

Type A errors include replacements which change the semantic meaning of the sentence, introduce non-words into the sentence, introduce co-reference errors within the sentence, or introduce words of the wrong class (e.g. replacement of a noun with an adjective).

Type B errors consist of misspelled words, definite/indefinite article or modifier errors, and erroneously inflected words.

The two types of errors can be viewed in terms of severity. Type B errors are generally the result of inaccuracies in the underlying dependencies on the Granska Tagger, or simply a matter of not compensating for the need to change articles as a result of a substitution. The majority of these errors could be managed by increasing the precision of the inflection handler, and by handling changes in articles iteratively, by changing the inflection of dependent words. This lies outside the scope of this thesis. The type B errors are considered *mild* in the sense that they are not in themselves the result of the strategy used for synonym replacement in this study. Type A errors on the other hand are considered *severe*. These errors are generally the result of the strategy employed by the replacement module and are relevant to estimating the performance of the modules.

The distinction between type A and type B errors require that the manual employed by the rater is strict enough to protect against rater bias. In order to verify that the manual's definition of errors was sufficient the inter-rater reliability was tested.

4.3 Inter-rater reliability

A pseudo-randomised portion of modified texts were used to test inter-rater reliability. The texts were modified without thresholds using word length or word frequency as the strategy for synonym replacement. The texts were divided evenly between the replacement modules which employed the inflection handler in half of the texts being evaluated. The texts were balanced across the modules based on genre, so that each module modified one text from each genre. The independent rater had no knowledge of which module had generated which text, and was not informed about the techniques employed by the different modules.

The inter-rater reliability was evaluated as the number of disagreements between the independent rater and the author, divided by the total number of replacements, that is, the maximum number of possible disagreements[1].

[1] *Cohen's Kappa* could be used to estimate the inter-rater reliability using the proportion of chance agreement for the three categories, *type A error*, *type B error*,

The average proportion of agreement between the two raters was 91.3%. For the four separate genres the average agreement was higher, accept for the FOKASS-texts which received an average of 85.5% agreement. The reason for the lower rate of agreement in this genre lies in that some terminology is repeated throughout the texts and disagreement between raters on the replacement of one term would often propagate throughout the whole text. For example, in one text the replacement of the word *tillfällig* (temporary) with *momentan* (momentary) gave rise to approximately one third of all disagreements. Tables 4.2–4.6 show the agreement percentages for all texts, and the genres respectively.

Table 4.2: Total proportion of inter-rater agreement for all texts.

Agreement	%
Type A	93.2
Type B	99.0
Total average	91.3

Table 4.3: Proportion of inter-rater agreement for ACADEMIC.

Agreement	%
Type A	95.7
Type B	99.0
Total average	93.7

Table 4.4: Proportion of inter-rater agreement for FOKASS.

Agreement	%
Type A	89.2
Type B	98.2
Total average	85.5

and *no error*, however, this would add very little to this study given that proportion of agreement between the two raters in this three-choice-task is high.

Table 4.5: Proportion of inter-rater agreement for FOF.

Agreement	%
Type A	92.9
Type B	99.7
Total average	92.3

Table 4.6: Proportion of inter-rater agreement for DN.

Agreement	%
Type A	95.2
Type B	99.6
Total average	94.3

Based on this cross section of inter-rater validated disagreements the manual was updated to handle previously diffuse descriptions of errors. One such change was the inclusion of "spoken language equivalents" as correct replacements for words, e.g. *va* (the common pronunciation) can be a correct replacement of *vad* (the correct spelling). The manual for analysis of errors was further updated by clarifying the instances in which the substitution of terminology should be approved. The initial validations of all modified texts were then updated according to the modified manual (see Appendix A).

4.4 Creating answer sheets

As a result of the inter-rater reliability test it was noted that the error analysis by hand was in need of some assistance. Not only is the method of manually marking up words with types of errors, and summarising errors and replacements, in a document very time consuming but it is also difficult to cross-check modifications of a text to validate that the texts have been judged using the same criteria. For this purpose a program was developed that allowed the rater to mark up the errors in a human readable fashion after which the results could be stored away in a more formal fashion. The program allowed the rater to modify previously defined an-

swer sheets by opening the document in question, loading its previously created answer sheet, and then updating it accordingly. Figure 4.1 shows the visual layout of the program. Loaded text files were automatically split up into sentences and words. Replaced words were marked up with the symbols '<' and '>' in the synonym replacement modules and only these words could be marked up as errors. Errors were entered in to the program by simply clicking a word repeatedly, marking it up as a correct replacement, type B error, or a type A error. The colors green, yellow, and red were used to visually distinguish the status of a replacement.

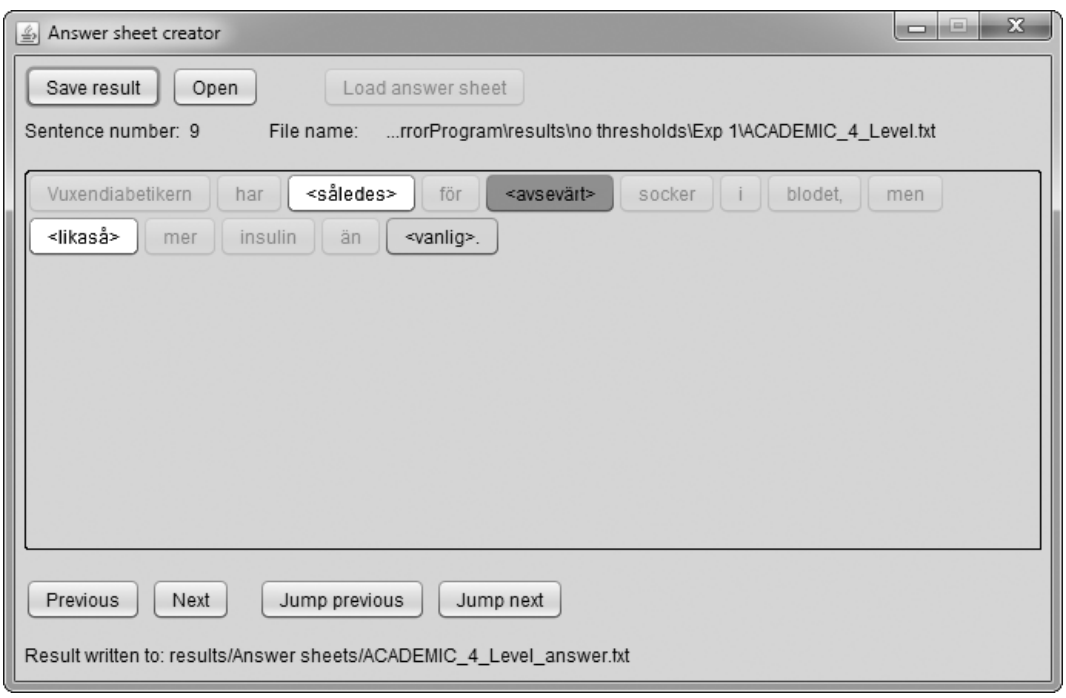

Figure 4.1: The graphical layout of the program used to create and edit answer sheets for the modified documents. In the example the original sentence "Vuxendiabetikern har därför för mycket socker i blodet, men också mer insulin än normalt" has been replaced by "Vuxendiabetikern har *således* för *avsevärt* socker i blodet, men *likaså* mer insulin än *vanlig*". Two errors have been marked up: *avsevärt* as a type A error (dark grey), and *vanlig* as a type B error (light grey). The rater could use the buttons previous or next to switch between sentences, or choose to jump to the next or previous sentence containing at least one replaced word.

4.5 Description of experiments

The 16 texts were processed using the different synonym replacement modules based on word frequency, word length, and level of synonymy. Word frequency as a criteria for replacement was motivated by the idea that word frequency can function as an estimate of reader familiarity with a word. Word length, on the other hand, was motivated by the various readability metrics which have shown that readability correlates with word length, that is, as the readability of a text increases the average word length decreases. The level of synonymy was used to estimate the accuracy of the SynLex synonym dictionary. Given that each synonym pair contains an estimate of synonym strength, 3.0–5.0, where 5.0 corresponds to the strongest synonym pairs, it is of interest to test whether a threshold can be introduced that maximizes the amount of replacements while simultaneously minimizing the amount of errors.

In the study type B errors are considered mild (see section 4.2) and will be ignored in the analysis of the results. These errors are almost exclusively a result of the imperfections in the Granska Tagger lemmatizer and the inflection handler, none of which are the target of analyzis in this study. The ratio of type A errors per replacement is therefore used to estimate the precision of the replacement modules.

The following sections describes the four experiments that were run in this study.

4.5.1 Experiment 1

Synonym replacement was performed on the 16 texts using a one-to-one matching between the words in the original text and the words in the synonym list. Since the inflection handler was not included only words written in their lemma form were evaluated for substitution.

4.5.2 Experiment 2

In experiment 2 the inflection handler was introduced. Its function was twofold: (1) synonym replacement takes place at the lemma level which dramatically increases the amount of words considered for replacement, and (2) it functions as an extra filter for the synonym replacements, since only words that have an inflection form corresponding to that of the word being replaced is allowed to be used as a replacement.

4.5.3 Experiment 3

In experiment 3 thresholds were introduced. The thresholds were incrementally increased and the generated texts were analyzed for errors in order to check for relationships between the level at which a replacement word was accepted and the error ratio. Since all replacements run the risk of introducing an error of type A the benefit of a replacement should be viewed in relation to the affect it has on the readability of the text. Using the templates created for the replacements the analyzis of errors could be performed automatically for each change in threshold.

4.5.4 Experiment 4

In experiment 4 the interaction effects of the strategies were studied. Investigating the entire spectrum of possible interaction effects at various threshold levels is not feasible in this study, given that in all instances where replacements are unpredictable a manual analyzis of errors must be performed. Instead only word frequency, which has the strongest support in research literature, was combined with level of synonymy. The motivation for the synonym replacement using word frequency was that the alternative word should be sufficiently more familiar than the original word in order to be considered simpler. The frequency threshold was set to 2.0, meaning that only replacement words with a frequency count of more than two times that of the original word was accepted. At the same time the threshold for the minimum level of synonymy of the alternative word was set to 4.0 in order to ensure that the quality of the synonym would be high.

If a word has more than one synonymous word that meets the requirements for replacement it can be argued that either the most frequent word, which is likely to be the most simple word, or the word with the highest level of synonymy, which is more likely to be a correct synonym, should be chosen. In experiment 4 both of these alternatives were investigated.

Chapter 5

Results

This chapter presents the results of the experiments that were run in this study. The modules on which the experiments were run are described in Chapter 3.

5.1 Experiment 1: Synonym replacement

This section presents the results from experiment 1 described in section 4.5. For more information about the modules used in this experiment see Chapter 3.

5.1.1 Synonym replacement based on word frequency

The results presented in Table 5.1 show that the replacement strategy based on word frequency resulted in an improvement in all readability metrics for every genre, and for the texts in general.

The greatest decrease in LIX-value was by 1.6 points (FOKASS), while the smallest decrease was by 1.2 points (FOF). The average decrease for all texts was by 1.4 points. The greatest decrease in OVIX-value was by 2.2 points (FOF), while the smallest decrease was by 0.8 points (FOKASS). The average decrease for all texts was by 1.5 points. The greatest decrease in proportion of long words, that is, words of six characters or more, was by 1.5% (FOKASS), and the smallest decrease was by 1.1% (FOF). The average decrease for all texts was 1.3%. Average word lengths decreased by 0–0.1 characters for all genres.

Table 5.1: Average LIX, OVIX, *proportion of long words* (LWP), and *average word length* (AWL) for synonym replacement based on word frequencies. Parenthesized numbers represent original text values. Bold text indicates that the change was significant compared to the original value.

Genre	LIX	OVIX	LWP (%)	AWL
ACADEMIC	51.5 (53.0)	**65.1** (66.5)	**27.2** (28.5)	**5.0** (5.1)
DN	**39.9** (41.3)	**65.4** (66.9)	**21.5** (22.7)	**4.7** (4.7)
FOF	43.3 (44.5)	**75.3** (77.5)	**25.7** (26.8)	**4.9** (5.0)
FOKASS	42.2 (43.8)	**48.3** (49.1)	**24.1** (25.6)	**5.1** (5.1)
All texts	**44.2** (45.6)	**63.5** (65.0)	**24.6** (25.9)	**4.9** (5.0)

The errors produced by the module is presented in Table 5.2. The results show that that the amount of erroneous replacements is very high, on average more than half of all replacements have been marked as errors, 0.52. The number of errors per replacement is most severe for ACADEMIC and FOF, 0.59, and best for DN, 0.43. A one-way ANOVA was used to test for differences among the four categories of text in terms of error ratio, but there was no significant difference, $F(3, 12) = .59$, $p = .635$. The results indicate that error ratio is not dependent on text genre.

Table 5.2: Average number of type A errors, replacements, and error ratio for replacement based on word frequency. Standard deviations are presented within brackets.

Genre	Errors (%)	Replacements	Error ratio
ACADEMIC	37.5 (18.7)	67.3 (15.8)	.59 (.36)
DN	16.3 (7.6)	36.5 (11.2)	.43 (.16)
FOF	27.0 (16.1)	46.3 (26.7)	.59 (.13)
FOKASS	26.3 (14.7)	56.0 (18.5)	.45 (.14)
All texts	26.8 (15.4)	51.5 (20.6)	.52 (.21)

5.1.2 Synonym replacement based on word length

The results presented in Table 5.3 show that the replacement strategy based on word length resulted in an improvement in terms of readability for every genre, and for the texts in general, in all readability metrics.

The greatest decrease in LIX-value was by 4.3 points (ACADEMIC), while the smallest decrease was by 3.1 points (DN). The average decrease for all texts was by 3.7 points. The greatest decrease in OVIX-value was by 1.3 points (DN and FOF), while the smallest decrease was by 0.7 points (FOKASS). The average decrease for all texts was 1.0 points. The greatest decrease in proportion of long words was by 3.8% (ACADEMIC and FOKASS), and the smallest decrease was by 2.7% (DN). The average decrease for all texts was 3.4%. Average word length decreased by 0.2 characters for all genres.

Table 5.3: Average LIX, OVIX, *proportion of long words* (LWP), and *average word length* (AWL) for synonym replacement based on word length with inflection handler. Parenthesized numbers represent original text values. Bold text indicates that the change was significant compared to the original value.

Genre	LIX	OVIX	LWP (%)	AWL
ACADEMIC	**48.7** (53.0)	**65.6** (66.5)	**24.7** (28.5)	**4.9** (5.1)
DN	**38.2** (41.3)	65.6 (66.9)	**20.0** (22.7)	**4.5** (4.7)
FOF	**41.1** (44.5)	76.2 (77.5)	**23.7** (26.8)	**4.8** (5.0)
FOKASS	**39.6** (43.8)	**48.4** (49.1)	**21.8** (25.6)	**4.9** (5.1)
All texts	**41.9** (45.6)	**64.0** (65.0)	**22.5** (25.9)	**4.8** (5.0)

The errors produced by the module is presented in Table 5.4. The results show that that the amount of erroneous replacements for this module is very high. The number of errors per replacement is worst for FOF, 0.71, and best for ACADEMIC, 0.52. The average error ratio was 0.59, that is, more than half of all words replaced were marked erroneous, and no genre had an error ratio below 50%. A one-way ANOVA was used to test for differences among four categories of text in terms of error ratio, but there was no significant difference, $F(3, 12) = 1.58$, $p = .245$. The results indicate that error ratio is not dependent on text genre.

Table 5.4: Average number of type A errors, replacements, and error ratio for replacement based on word length. Standard deviations are presented within brackets.

Genre	Errors	Replacements	Error ratio
ACADEMIC	51.5 (19.8)	103.3 (35.6)	.52 (.21)
DN	27.8 (3.3)	50.5 (10.1)	.57 (.13)
FOF	52.0 (34.6)	73.0 (49.7)	.71 (.08)
FOKASS	69.5 (13.8)	125.5 (12.2)	.55 (.06)
All genres	50.2 (24.3)	88.1 (40.9)	.59 (.14)

5.1.3 Synonym replacement based on level of synonymy

The readability metrics are less important for this module, since replacements are performed regardless of whether the new word is *easier* than the original. The results are however relevant as a reference in the discussion to follow.

The results in table 5.5 shows that for all genres the replacement based on level of synonymy affected the readability metrics negatively except for the OVIX-value. The greatest increase in LIX-value was by 2.9 points (DN), while the smallest increase was by 1.2 points (ACADEMIC). The average increase for all texts was by 2.1 points. The OVIX-value decreased by at most 0.2 points for all genres except DN for which it increased by 0.1 points. The greatest increase in proportion of long words was by 2.7% (DN), and the smallest increase was by 1.1% (ACADEMIC). The average increase for all texts was by 1.9%. Average word length increased by 0.2 characters for DN, and by 0.1 characters for the other genres.

Table 5.5: Average LIX, OVIX, *proportion of long words* (LWP), and *average word length* (AWL) for synonym replacement based on level of synonymy. Parenthesized numbers represent original text values. Bold text indicates that the change was significant compared to the original value.

Genre	LIX	OVIX	LWP (%)	AWL
ACADEMIC	**54.2** (53.0)	66.3 (66.5)	**29.6** (28.5)	5.2 (5.1)
DN	**44.2** (41.3)	67.0 (66.9)	**25.4** (22.7)	4.9 (4.7)
FOF	**47.2** (44.5)	77.3 (77.5)	**26.8** (29.2)	**5.1** (5.0)
FOKASS	45.3 (43.8)	48.9 (49.1)	27.0 (25.6)	**5.2** (5.1)
All texts	**47.7** (45.6)	64.9 (65.0)	**27.8** (25.9)	**5.1** (5.0)

The errors produced by the module is presented in Table 5.6. The results show that that the amount of erroneous replacements is high. The number of errors per replacement is highest for DN, 0.56, and best for FOKASS, 0.45. A one-way ANOVA was used to test for differences among four categories of texts in terms of error ratio, but there was no significant difference, $F(3, 12) = 2.15$, $p = .147$. The results indicate that error ratio is not dependent on text genre.

Table 5.6: Average number of type A errors, replacements, and error ratio for replacement based on level of synonymy. Standard deviations are presented within brackets.

Genre	Errors	Replacements	Error ratio
ACADEMIC	87.5 (32.5)	181.8 (62.1)	.48 (.08)
DN	66.5 (16.6)	117.5 (19.2)	.56 (.05)
FOF	82.3 (56.2)	150.8 (87.8)	.53 (.09)
FOKASS	99.8 (15.1)	222.0 (31.3)	.45 (.03)
All genres	84.0 (33.1)	168.0 (64.6)	.50 (.08)

5.2 Experiment 2: Synonym replacement with inflection handler

This section presents the results from experiment 2 described in section 4.5. For more information about the modules used in this experiment see Chapter 3.

5.2.1 Synonym replacement based on word frequency

The results presented in Table 5.7 show that the replacement strategy based on word frequency resulted in an improvement in terms of readability for every genre, and for the texts in general, in all readability metrics. The greatest decrease in LIX-value was by 2.4 points (FOKASS), while the smallest decrease was by 0.9 points (ACADEMIC). The average decrease for all texts was by 1.6 points. The greatest decrease in OVIX-value was by 1.9 points (ACADEMIC), while the smallest decrease was by 0.8 points (FOKASS). The average decrease for all texts was by 1.4 points. The greatest decrease in proportion of long words was by 2.1% (FOKASS), while there was an increase of 0.9% for DN. The average decrease for all texts was 1.5%. Average word lengths decreased by 0–0.1 characters for all genres.

Table 5.7: Average LIX, OVIX, *proportion of long words* (LWP), and *average word length* (AWL) for synonym replacement based on word frequencies with inflection handler. Parenthesized numbers represent original text values. Bold text indicates that the change was significant compared to the original value.

Genre	LIX	OVIX	LWP (%)	AWL
ACADEMIC	**52.1** (53.0)	**64.6** (66.5)	27.8 (28.5)	**5.0** (5.1)
DN	40.0 (41.3)	**65.7** (66.9)	22.7 (21.8)	4.7 (4.7)
FOF	42.5 (44.5)	**75.8** (77.5)	24.8 (26.8)	**4.9** (5.0)
FOKASS	41.4 (43.8)	48.3 (49.1)	23.5 (25.6)	**5.0** (5.1)
All texts	**44.0** (45.6)	**63.6** (65.0)	**24.4** (25.9)	**4.9** (5.0)

The errors produced by the module is presented in Table 5.8. The results show that that the amount of erroneous replacements is high. The number of errors per replacement is most severe for ACADEMIC, 0.37,

and best for FOKASS, 0.31. A one-way ANOVA was used to test for differences among four categories of text in terms of error ratio, but there was no significant difference, $F(3, 12) = .43$, $p = .739$. The results indicate that error ratio is not dependent on text genre.

Table 5.8: Average number of type A errors, replacements, and error ratio for replacement based on word frequency with inflection handler. Standard deviations are presented within brackets.

Genre	Errors	Replacements	Error ratio
ACADEMIC	38.8 (4.9)	105.3 (9.6)	.37 (.04)
DN	17.3 (8.1)	52.3 (12.1)	.32 (.11)
FOF	26.5 (20.5)	70.3 (39.5)	.35 (.08)
FOKASS	19.3 (5.1)	67.3 (25.4)	.31 (.10)
All texts	25.4 (13.5)	73.8 (29.9)	.34 (.08)

5.2.2 Synonym replacement based on word length

The results presented in Table 5.9 show that the replacement strategy based on word length resulted in an improvement in terms of readability for every genre, and for the texts in general, in all readability metrics. The greatest decrease in LIX-value was by 6.1 points (ACADEMIC), while the smallest decrease was by 3.8 points (DN). The average decrease for all texts was by 5.1 points. The greatest decrease in OVIX-value was by 1.3 points (ACADEMIC), while the smallest decrease was by 0.4 points (FOKASS). The average decrease for all texts was 0.8 points. The greatest decrease in proportion of long words was by 5.2% (ACADEMIC), and the smallest decrease was by 3.2% (DN). The average decrease for all texts was 4.6%. Average word length decreased by 0.3 characters for ACADEMIC and FOF, and by 0.2 characters for DN and FOKASS.

Table 5.9: Average LIX, OVIX, *proportion of long words* (LWP), and *average word length* (AWL) for synonym replacement based on word length with inflection handler. Parenthesized numbers represent original text values. Bold text indicates that the change was significant compared to the original value.

Genre	LIX	OVIX	LWP (%)	AWL
ACADEMIC	**46.9** (53.0)	**65.2** (66.5)	**23.3** (28.5)	**4.8** (5.1)
DN	**37.5** (41.3)	**66.1** (66.9)	**19.5** (22.7)	**4.5** (4.7)
FOF	**39.1** (44.5)	76.6 (77.5)	**22.0** (26.8)	**4.7** (5.0)
FOKASS	**38.3** (43.8)	48.7 (49.1)	**20.5** (25.6)	**4.9** (5.1)
All texts	**40.5** (45.6)	**64.2** (65.0)	**21.3** (25.9)	**4.7** (5.0)

The errors produced by the module is presented in Table 5.10. The results show that that the amount of erroneous replacements for this module is high. The number of errors per replacement is worst for FOKASS, 0.47, and best for ACADEMIC, 0.37. A one-way ANOVA was used to test for differences among four categories of text in terms of error ratio, but there was no significant difference, $F(3, 12) = 3.20$, $p = .062$. The results indicate that error ratio is not dependent on text genre.

Table 5.10: A number of type A errors, replacements, and error ratio for replacement based on word length with inflection handler. Standard deviations are presented within brackets.

Genre	Errors	Replacements	Error ratio
ACADEMIC	56.3 (15.0)	152.8 (38.9)	.37 (.04)
DN	24.5 (9.8)	61.0 (18.7)	.39 (.05)
FOF	48.8 (38.2)	99.0 (57.6)	.46 (.09)
FOKASS	54.2 (14.4)	115.8 (34.0)	.47 (.03)
All genres	45.9 (23.9)	107.1 (49.3)	.42 (.07)

5.2.3 Synonym replacement based on level of synonymy

The readability metrics are less important for this module, since replacements are performed regardless of whether the new word is *easier* than

the original. The results are, however, relevant as a reference in the discussion to follow. The results in table 5.11 shows that for all genres the replacement based on level of synonymy affected the readability metrics negatively for all genres and all metrics accept for the OVIX-value for ACADEMIC. The greatest increase in LIX-value was by 4.6 points (DN), while the smallest increase was by 1.8 points (FOKASS). The average increase for all texts was by 2.9 points. The greatest increase in OVIX-value was by 0.3 points (DN), while the only decrease was by 0.6 points (ACADEMIC). The greatest increase in proportion of long words was by 4.1% (DN), and the smallest increase was by 1.7% (FOKASS). The average increase for all texts was by 2.5%. Average word length increased by at most 0.3 characters (DN), and the smallest increase was by 0.1 characters (FOKASS). The average increase in average word length was by 0.2 characters.

Table 5.11: Average LIX, OVIX, *proportion of long words* (LWP), and *average word length* (AWL) for synonym replacement based on level of synonymy with inflection handler. Parenthesized numbers represent original text values. Bold text indicates that the change was significant compared to the original value.

Genre	LIX	OVIX	LWP (%)	AWL
ACADEMIC	**55.4** (53.0)	**65.9** (66.5)	30.5 (28.5)	5.2 (5.1)
DN	**45.9** (41.3)	67.1 (66.9)	**26.8** (22.7)	**5.0** (4.7)
FOF	**47.2** (44.5)	77.8 (77.5)	**29.3** (26.8)	**5.2** (5.0)
FOKASS	45.6 (43.8)	49.2 (49.1)	27.3 (25.6)	5.3 (5.1)
All texts	**48.5** (45.6)	65.0 (65.0)	**28.4** (25.9)	**5.2** (5.0)

The errors produced by the module is presented in Table 5.12. The results show that that the amount of erroneous replacements is high. The number of errors per replacement is most severe for ACADEMIC, 0.46, and best for DN, 0.40. A one-way ANOVA was used to test for differences among four categories of text in terms of error ratio, but there was no significant difference, $F(3, 12) = 2.39$, $p = .120$. The results indicate that error ratio is not dependent on text genre.

Table 5.12: Average number of type A errors, replacements, and error ratio for replacement based on level of synonymy with inflection handler. Standard deviations are presented within brackets.

Genre	Errors	Replacements	Error ratio
ACADEMIC	134.5 (24.1)	290.3 (54.3)	.46 (.03)
DN	62.3 (12.7)	154.8 (30.1)	.40 (.04)
FOF	98.0 (57.7)	216.3 (57.7)	.44 (.03)
FOKASS	101.0 (13.7)	234.8 (49.6)	.44 (.03)
All genres	98.9 (39.4)	224.0 (80.3)	.44 (.04)

5.3 Experiment 3: Threshold estimation

This section presents the results from experiment 3 described in section 4.5. For more information about the modules used in this experiment see Chapter 3.

5.3.1 Synonym replacement based on word frequency

A threshold for replacements based on word frequency count was introduced and increased incrementally. Since the module makes replacements only with the synonyms of the highest frequency raising the threshold will exclude substitutions of words in a predictable fashion. Word counts vary a lot and rather than introducing a numeric threshold for an alternative word the threshold value was expressed relative to the original word's frequency count.

In the graph in Figure 5.1 the error ratio for the all texts for the thresholds between 1.0 and 30.0 is displayed (1.0 corresponds to the no-threshold replacement technique employed in experiment 2). The graph shows that there is no clear relationship between threshold and error ratio when viewed together. For some texts the error ratio decreases as the threshold increases, but for others the opposite is true. Clustering particularly occurs around values around the maximum values for the two variables.

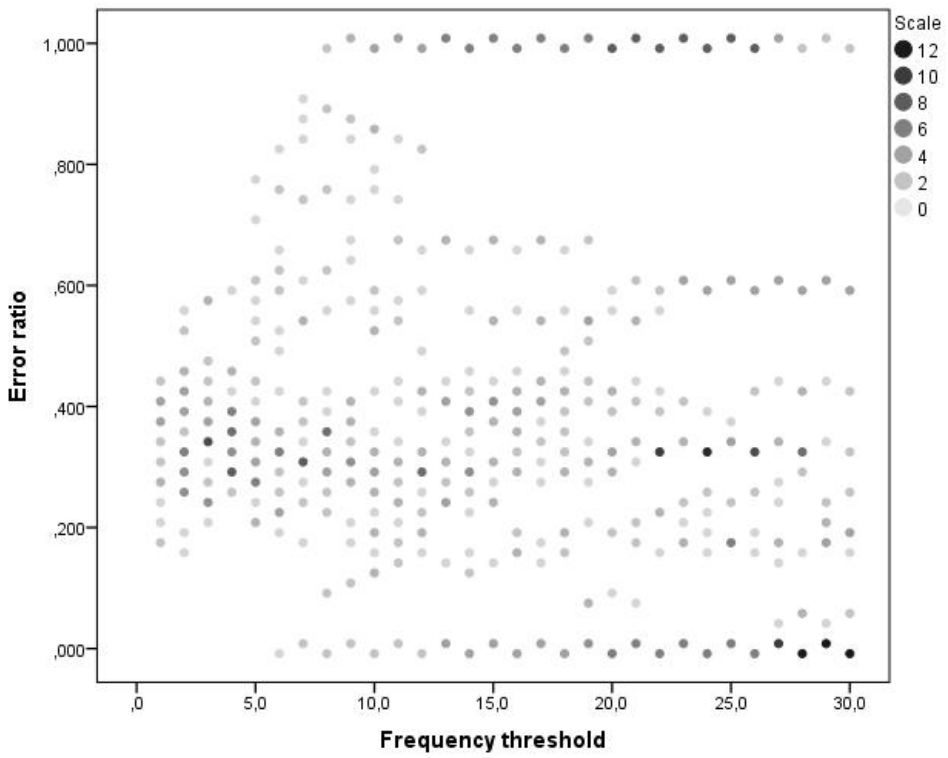

Figure 5.1: The error ratio in relation to frequency threshold for all texts. The opacity of the black dots indicates the amount of clustering around a coordinate, darker dots indicate a higher degree of clustering.

The graph in Figure 5.2 shows the threshold and error ratio summarized for the texts in their respective genres. A weak, but significant, correlation between threshold and error ratio exists for ACADEMIC, $r(234) = -.205$, $p < .01$, DN, $r(234) = -.231$, $p < .001$, and a positive correlation for FOF, $r(234) = .197$, $p < .01$. As before, the result of this experiment depends almost exclusively on the nature of individual text, rather than on which genre it belongs too.

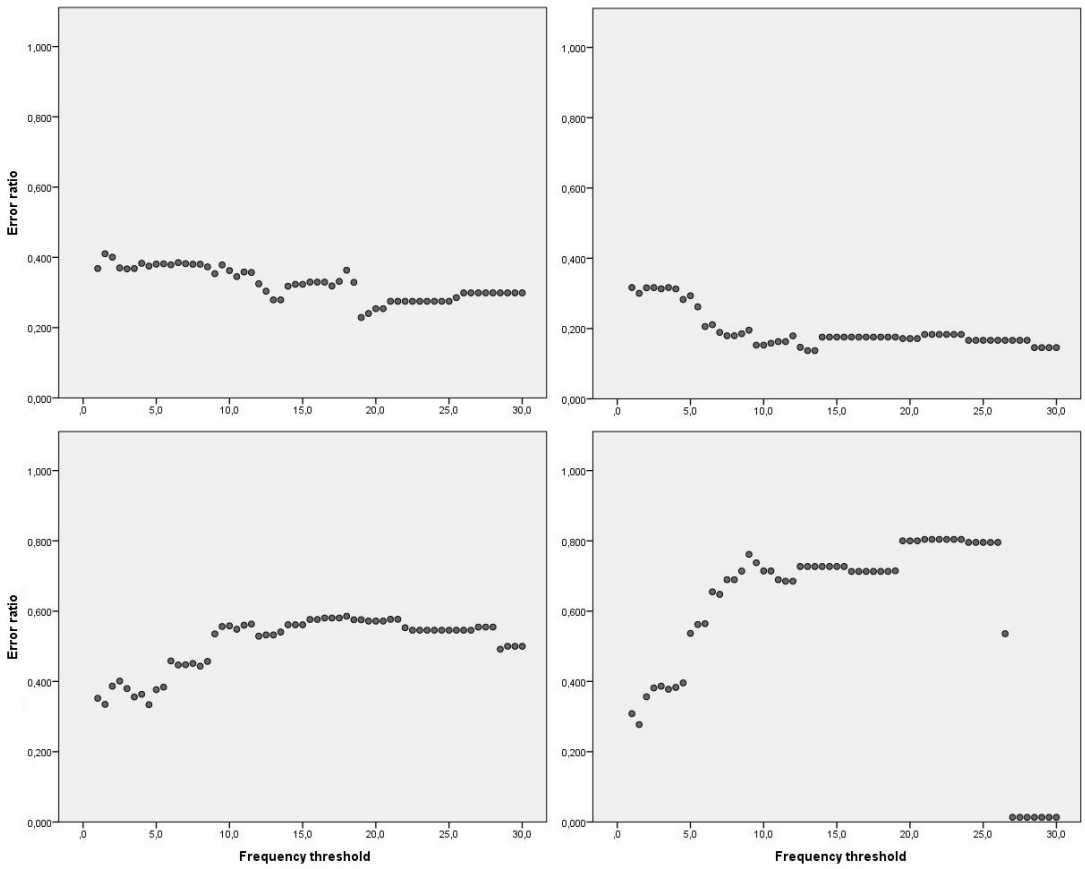

Figure 5.2: The error ratio in relation to frequency threshold for summarized values for genres: ACADEMIC (top left), DN (top right), FOF (lower left), and FOKASS (lower right).

5.3.2 Synonym replacement based on word length

A threshold for replacements based on word length was introduced and increased incrementally by one character at a time. The module makes replacements only with the shortest synonyms and the threshold will exclude substitutions of words in a predictable fashion. In the graph in Figure 5.5 the error ratio for the all texts for the thresholds 0–7 characters is displayed (0 corresponds to the no-threshold replacement technique employed in experiment 2). The graph shows that there is no clear relationship between threshold and error ratio when viewed together.

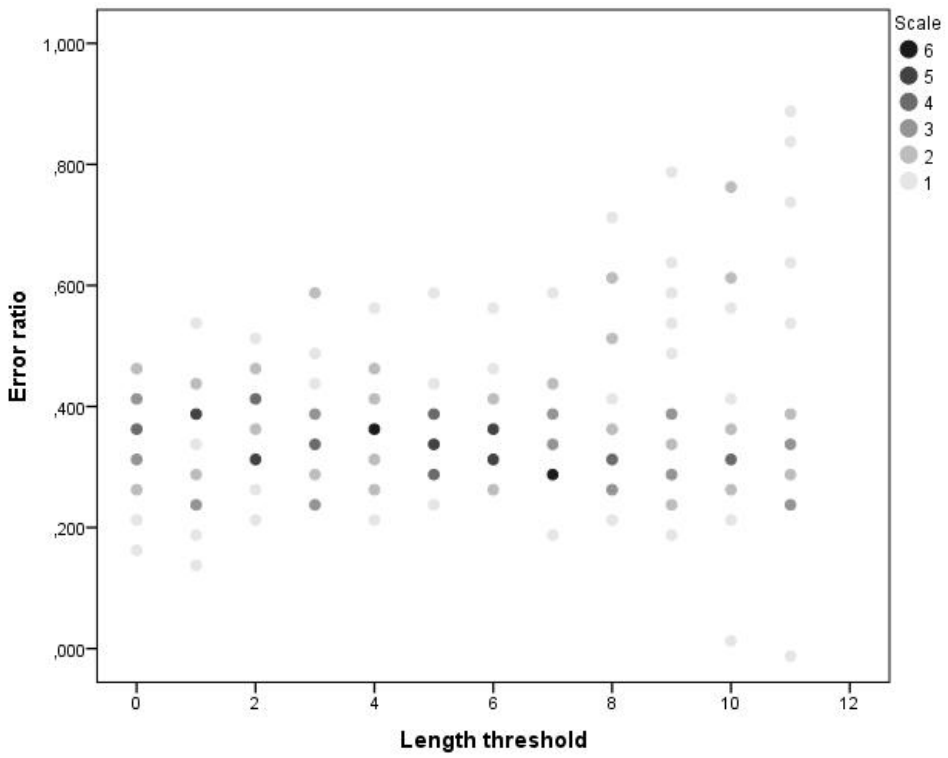

Figure 5.3: The error ratio in relation to length threshold for all texts. The opacity of the black dots indicates the amount of clustering around a coordinate, darker dots indicate a higher degree of clustering.

The graph in Figure 5.4 shows the threshold and error ratio summarized for the texts in their respective genres. A significant correlation between threshold and error ratio exists for DN, $r(46) = -.336$, $p < .05$, and FOKASS, $r(46) = -.661$, $p < .001$. As before, in general the results depend almost exclusively on the nature of individual text, rather than on which genre it belongs too, only in FOKASS the relationship was strong.

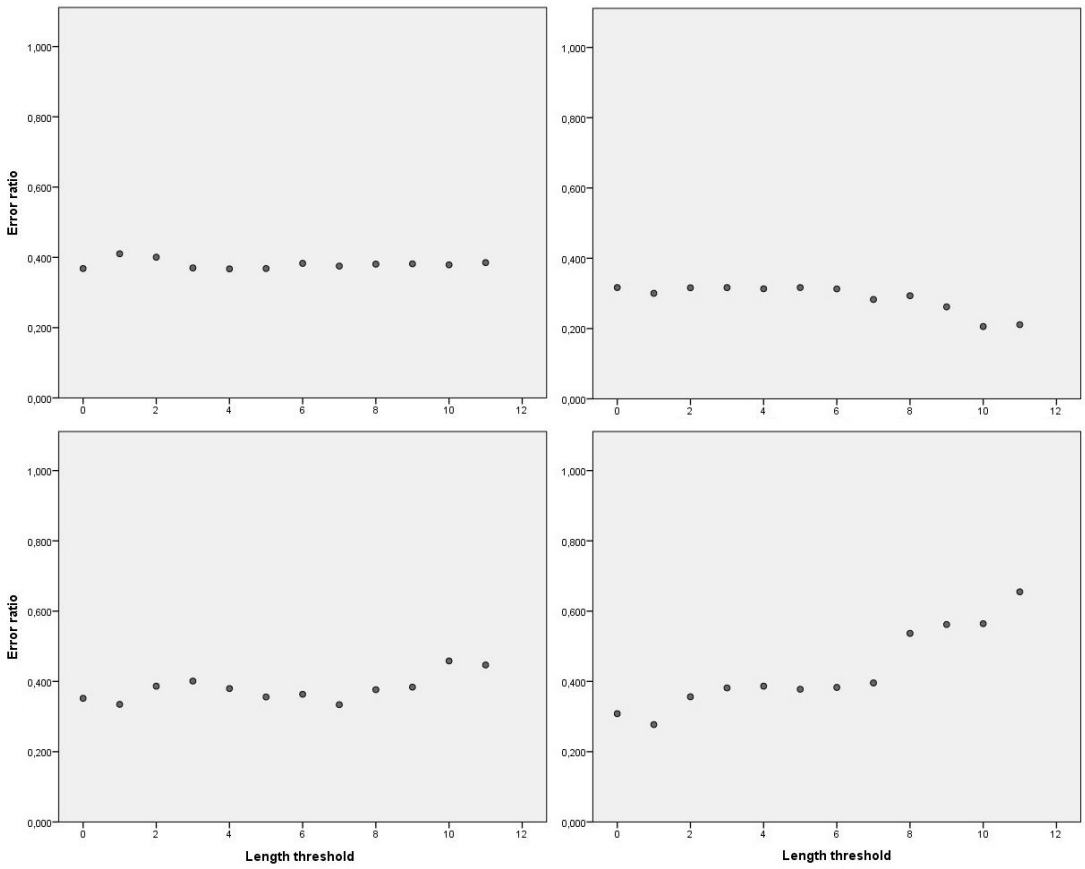

Figure 5.4: The error ratio in relation to length threshold for summarized values for genres: ACADEMIC (top left), DN (top right), FOF (lower left), and FOKASS (lower right).

5.3.3 Synonym replacement based on level of synonymy

A threshold for replacements based on level of synonymy was introduced and increased incrementally by 0.1 points. The module makes replacements only with the synonyms of highest level of synonymy and the threshold excludes substitutions of words in a predictable fashion, removing first those replacements with weak synonymy level. In the graph in Figure 5.3 the error ratio for the all texts for the thresholds from 3.0 to 5.0 (3.0 corresponds to the no-threshold replacement technique employed in experiment 2). The graph shows that there is no clear relationship between threshold and error ratio when viewed together.

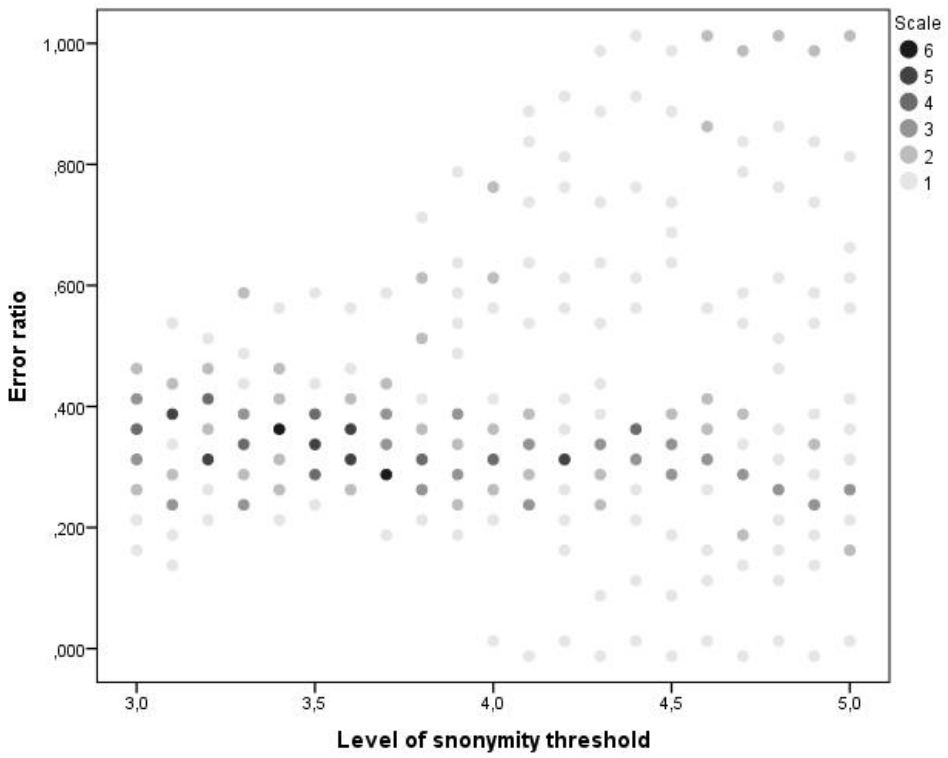

Figure 5.5: The error ratio in relation to level of synonymy threshold for all texts. The opacity of the black dots indicates the amount of clustering around a coordinate, darker dots indicate a higher degree of clustering.

The graph in Figure 5.6 shows the threshold and error ratio summarized for the texts in their respective genres. A significant negative correlation between threshold and error ratio exists for DN, $r(82) = -.498$, $p < .001$, and a positive correlation exists for FOF, $r(46) = .370$, $p < .001$, and FOKASS, $r(46) = .607$, $p < .001$. The results depend highly on the nature of individual text, rather than on which genre it belongs too.

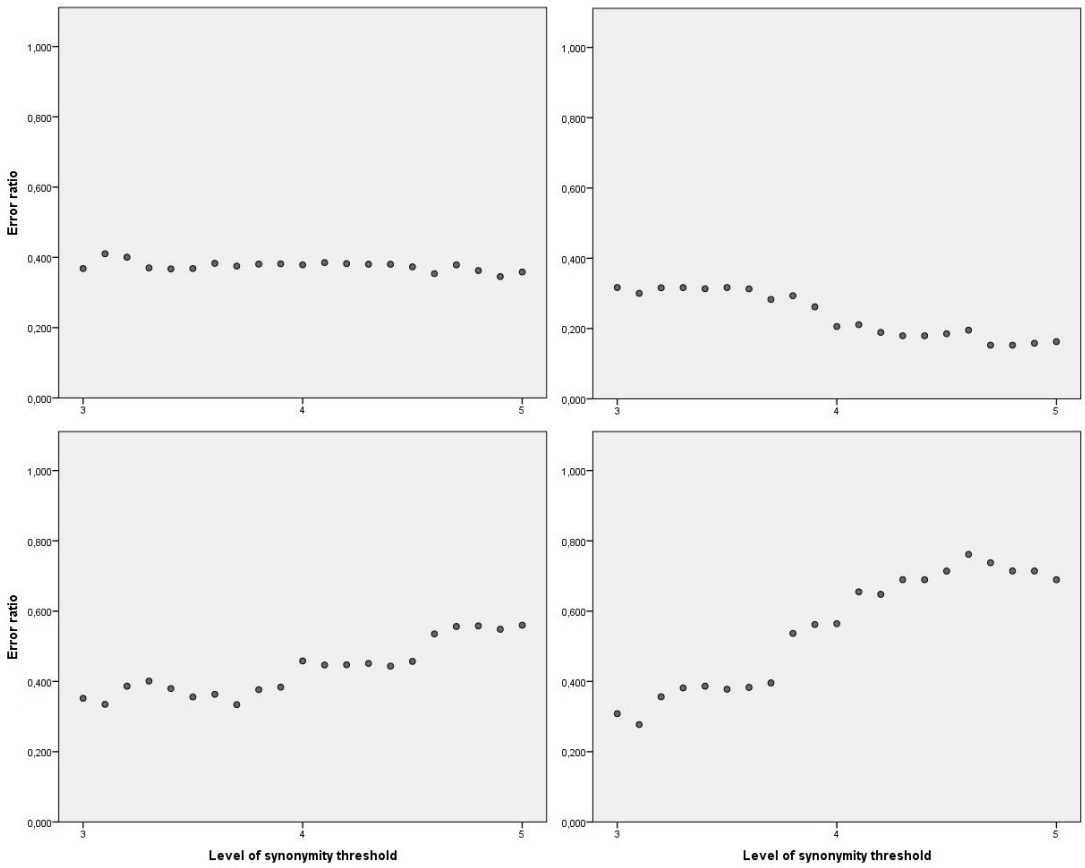

Figure 5.6: The error ratio in relation to level of synonymy threshold for summarized values for genres: ACADEMIC (top left), DN (top right), FOF (lower left), and FOKASS (lower right).

5.4 Experiment 4: Frequency combined with level of synonymy

In experiment 4 the interaction affects of word frequency strategy and level of synonymy was investigated. Predefined thresholds were used. For more information about the experiment see section 4.5.

A paired samples t-test was used to compare the performance of combining frequency and level of synonymy with frequency alone (Freq), which was the best performing strategy from experiment 2. The threshold for frequency was set to 2.0 and the threshold for level of synonymy was set to 4.0. The experiment was run twice prioritizing either frequency (Pri-

oFreq) or level of synonymy (PrioLevel) when more than one synonym passed the thresholds. The words that are replaced are the always the same for both FreqPrio and LevelPrio, only the words used as replacements sometimes differ. Comparing the performance of the two strategies revealed no siginificant differences in terms of error ratio when comparing all text or the genres separately. The average number of replacements per text was less than one-fourth of the number of replacements performed by Freq, 8.0 compared to 34.0.

LevelPrio performed significantly better than Freq when considering all texts, $t(15) = 2.46$, $p < .05$. When comparing performance for the separate text genres LevelPrio performed significantly better than Freq only for DN, $t(3) = -4.69$, $p < .05$. For the other genres the difference was not significant, $t(3) = -.44$, $p = .69$ (ACADEMIC), $t(3) = -.76$, $p = .50$ (FOF), and $t(3) = -1.87$, $p = .16$ (FOKASS).

FreqPrio did not perform significantly better than Freq when considering all texts, $t(15) = 2.05$, $p = .06$. When looking at the separate text genres FreqPrio performed significantly better than Freq only for DN, $t(3) = -3.19$, $p < .05$. For the other genres the difference was not significant, $t(3) = -.17$, $p = .87$ (ACADEMIC), $t(3) = -.37$, $p = .74$ (FOF), and $t(3) = -1.75$, $p = .18$ (FOKASS).

Figure 5.7 shows the average performance of the different genres with error bars representing one standard deviation. The average error ratio for all texts was 0.27 for both FreqPrio and LevelPrio.

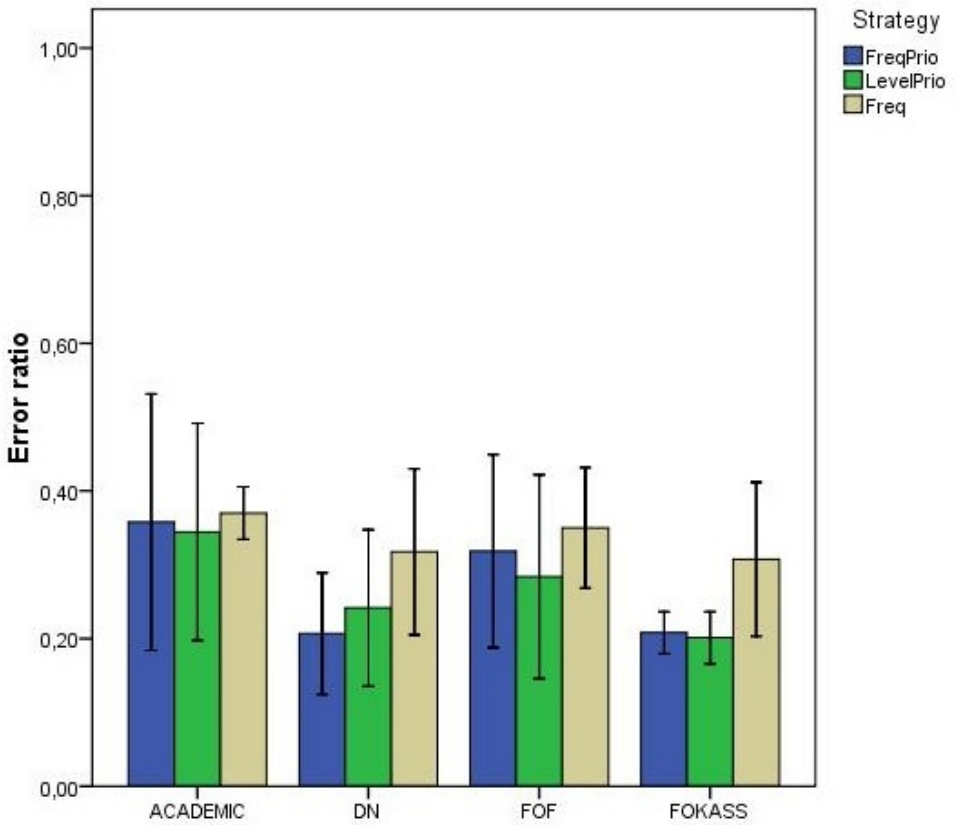

Figure 5.7: Average error ratio for replacements using 2.0 as threshold for frequency and 4.0 as threshold for level prioritizing frequency (FreqPrio), or level (LevelPrio), and the error ratio for replacements based on frequency only. Error bars represent one standard deviation.

Chapter 6

Analysis of results

This chapter discusses the results of the experiments presented in Chapter5. If no mention of the type of error is made explicitly error refers to type A errors.

6.1 Experiment 1

Experiment 1 evaluated the synonym dictionary in a very direct way using the three replacement strategies frequency count (FREQ), word length (LENGTH), and level of synonymy (LEVEL). Only open word classes written in lemma form were replaced.

6.1.1 FREQ

FREQ resulted in an overall improvement in terms of the readability metrics employed. The two most popular metrics, LIX and OVIX, both decreased for all genres. LIX is only affected by the replacements through the number of long words, which is the only LIX feature that is affected by one-to-one word replacements, but as it is a more familiar readability metric than average world length and proportion of long words it will be used when possible.

Though the decrease was small, on average 1.4 points for LIX and 1.5 points for OVIX, the results confirm two important assumptions of this thesis. Firstly, replacing words with the most common synonym results in an overall decrease of long words and an overall decrease in word

length. Though many popular readability metrics rely on word length as an estimation of word difficulty it is seldom discussed why this relationship exists. These results support theories that suggest that the length of words decreases with frequency of use. Readability metrics employing word length as a feature may actually indirectly be measuring word familiarity. Secondly, using a fixed strategy of this type replaces synonyms in such a way that the variation of words decreases, thus possibly decreasing the vocabulary load of the text.

The average error ratio for replacements based on frequency count was 52%, which in practice means that about half of all replacements performed resulted in a type A error. Even replacements in the most successful genre (FOKASS), in terms of error ratio, 45% of the substituted words resulted in errors. But if we look at the standard deviations of the separate genres in Table5.2 on page 30 we see that there is a great amount of individual variation within the genres. For ACADEMIC, where the performance was the lowest, an average of 59% of the words replaced were type A errors but the standard deviation was 36%. For the texts in general the standard deviation is considerably smaller, 21%, but still quite high.

6.1.2 LENGTH

LENGTH resulted in an improvement in terms of all readability metrics employed. The decrease in overall word length is a natural effect of the replacements since only shorter words are allowed to be used for substitution. This fixed strategy for replacing words result in smaller variation of the words used, potentially decreasing the overall vocabulary load of the text.

The average decrease in LIX-value was 3.7 points, which is more than twice as high as the decrease for FREQ. Interestingly the OVIX-value decreased by an average of only 1.0 points, which was less than for FREQ.

Turning to the ratio of errors we see that LENGTH on average performs worse than FREQ. For the texts in general the error ratio is 59% compared to 52% for FREQ. The standard deviation for error ratio in LENGTH, as seen in Table5.4 on page 32, was smaller for LENGTH than for FREQ. Synonym replacement in FOF produced a particularly high ratio of errors, 71%. At the same time replacements based on length also resulted in considerably higher average amount of replacements per texts, 88.1 replacements in LENGTH compared to 51.1 replacements in FREQ.

6.1.3 LEVEL

The substitutions made using the synonymy level strategy are interesting since they reflect the maximum number of possible replacements in the texts, regardless of whether the replacements can be motivated or not. The estimated effects of the substitutions on the readability metrics are not apparent from a theoretical perspective, and replacements are likely not affecting the overall readability metrics in any considerable way. In some ways LEVEL also partly reflects the precision of the SynLex dictionary since the strongest synonym available will be used for the replacement regardless of word frequency or word length.

The results of the experiment showed that the readability metrics were affected negatively by the replacements based only on level of synonymy. LIX increased by on average 2.1 points, and the average OVIX value decreased marginally. Average word length and proportion of long word increased somewhat.

The average ratio of errors produced by LEVEL for all texts was 50.0%, that is, using the best synonym available for every word in a text results in erroneous replacements for on average every second word. The number of replacements made by LEVEL is very high, on average 168.0 words are replaced in every text, which is more than twice the amount of replacements made using LENGTH.

6.2 Experiment 2

In experiment 2 the inflection handler was introduced. Its function was to inflect the alternative words before replacements were made. This indirectly works as a filter for words by requiring that the alternative word must be of the same word class [1] and that it can be inflected appropriately. The inflection also increases the amount of possible replacements since the comparisons are made on the level of lemmas. This affects the strategy employed in LENGTH since comparisons are made on the level of the lemma rather than the on the length of the inflected word.

For FREQ the replacements again resulted in an improvement in terms of readability metrics. The LIX-value remained relatively unchanged compared to the result in experiment 1, decreasing by on average 1.6 points. The OVIX-value decreased by 1.4 points, slightly less than in experiment 1. However, the average amount of replacements increased

[1] The synonym pairs in SynLex are not guaranteed to belong to the same word class.

markedly from 51.1 replacements in experiment 1 to 73.8 replacements using the inflection handler. In terms of error ratio there is a considerable improvement, from an average of 52% errors to an average of 34%. This value is probably, however, still too high to be applicable in a completely automatic simplification system.

When combining LENGTH with the inflection handler the average LIX-value drops by 5.1 points, which is actually a considerable drop in terms of estimated readability. For OVIX the decrease was smaller than in experiment 1. The number of replacements increased from 88.1 in experiment 1 to 107.1. The average error ratio also improved considerably, from an average error ratio of 59% in experiment 1 to 42%.

Combining level of synonymy with the inflection handler affected readability metrics negatively compared to the original text. LIX increased by 1.9 points, the other readability metrics remained about the same as in experiment 1. However, there was again a great increase in number of replacements, from an average of 168.0 to 224.0 replacements. At the same time the error ratio dropped from 50% to 44%.

6.3 Summary of experiment 1 and 2

Replacement of words with synonyms based on word frequency or word length results in an overall improvement with regard to the readability metrics LIX, OVIX, average word length, and proportion of long words. Including an inflection handler improves performance in terms of the readability metrics, number of words replaced, and average error ratio. There was, however, no significant difference in terms of error ratio between the four text genres for any of the three strategies employed.

In experiment 1 synonym replacement based on level of synonymy was the strategy with the best error ratio, but after the introduction of the inflection handler it performs worse than the other two strategies, despite an improved error ratio. The best performing strategy is replacement based on word frequencywith inflection handler, with an error ratio of 34%, but it still performs too poorly to be used in a non-supervised simplification system.

6.4 Analysis of experiment 3

In experiment 3 thresholds were introduced to the modules from experiment 2. Manipulation of the thresholds, paired with automatic error analysis based on the answer sheets created for experiment 2, was used to explore whether there existed some threshold level at which point the error ratio improved for the texts in general, or for a specific genre.

The frequency threshold was incrementally increased up to the point where the alternative word had to be at least 30 times more common than the original word in order for a replacement to occur. For each increase in threshold the error ratio was calculated. There was no significant correlation between the two variables, that is, synonyms that occur very frequently are roughly as probable to be correct substitutions as any other synonyms, given that they are at least as common as the original word. For the separate genres there were significant correlations, but these were sometimes positive and sometimes negative. It is likely that increasing the threshold for word frequency increases the average word familiarity for the words that are used as replacements, but asserting the relationship between frequency of occurrence for words and word familiarity lies outside the scope of this thesis.

A threshold was also introduced for LENGTH. The threshold for the required difference between the length of the original word and the replacement word was increased incrementally by one character until no substitutions were performed. For every threshold level the texts were automatically evaluated for errors based on its answer sheet from experiment 2. Again, no correlation existed for the texts in general, but for the genres there where some significant but small correlations both positive and negative.

It seems natural to assume that increasing the threshold for replacement in terms of level of synonymy would increase proportion of correct substitutions, as the synonym pairs should come closer to being absolute synonyms (see section 2.3.1 on page 10). Introducing this threshold and incrementally increasing the requirement for the replacements, however, had no major effect on error ratio for the texts in general. For FOKASS there was a significant correlation between level of synonymy and threshold, that is, as the threshold increased the ratio of errors decreased, aside from that there were no significant correlations.

6.5 Analysis of experiment 4

The lack of clear relationships between any of the thresholds in experiment 3 is intriguing, especially in the case of a threshold for level of synonymy. The absence of a strong connection between these thresholds does not exclude the possibility of interaction between the different strategies. Manipulating two or three thresholds at the same time would be desirable in order to evaluate all possible combinations of thresholds; however, this is not feasible for this study since the possible interaction effects of the strategies make replacements impossible to predict. This means that the evaluation of errors cannot be made automatically using the strategy employed in this study.

The most promising strategy for replacement in experiment 2 was found to be the one based on word frequency. This is also the strategy with the strongest theoretical bearing, and as commonly used words tend to become shorter over time, following Zipf's law (Dell'Orletta et al., 2011), it also affected readability metrics in a positive way. In experiment 4 fixed thresholds for frequency and level of synonymy were introduced. The combination of thresholds could be made in numerous ways but in this experiment thresholds were set to 2.0 for frequency and 4.0 for level of synonymy. If more than one word met the criteria words were either selected based on highest frequency or highest level of synonymy. This means that the two strategies employed in this experiment always replace the same words but that the replacement word may differ. When looking at the average error ratio of the composite strategies, 27% for both, it appears that they are superior to the previous strategies in terms error ratio, however, this difference was only significant in a very limited way, possibly due to large variations in the results. Prioritizing level of synonymy resulted in a reduced error ratio for the texts in general compared to replacing words based on frequency alone. Prioritizing frequency was, however, not significantly better than frequency alone for the texts in general. Both composite strategies resulted in an improvement in terms of error ratio for the genre DN, while the result for the other genres was not significant. It is important to view this result with regard to the number of replacements performed. The number of replacements go down as an effect of introducing thresholds, and as a result the effect on the readability metrics become less prominent with higher thresholds.

Chapter 7

Discussion

This chapter discusses the results and the implications of this study in relation to modern research, and elaborates on the limitations of the strategies that were employed in this study. Furthermore, suggestions on possible directions for improvement of automatic lexical simplification are discussed.

7.1 Limitations of the replacement strategies

The strategies employed in this study take into consideration the context in which the words occur only in an indirect and limited fashion, namely by considering part-of-speech tags that contain word class and word inflection information. The results of the experiments performed in this study demonstrate that the precision of these simple strategies is inadequate. Even when employing the maximum threshold criteria for level of synonymy, 5.0 in SynLex, the precision of replaced words is quite poor, about 80% correct replacements for the best of the genres, DN (newspaper texts), and much lower overall. Combining the two most promising strategies, word frequency and level of synonymy, which was done in experiment 4, only slightly improves the overall error ratio and seems to hold no real hope of raising accuracy sufficiently.

7.1.1 The dictionary

Various improvements could be made the dictionary used in this study. One way could be to use specific dictionaries, adapted to certain topics and/or types of readers. If we imagine two poor readers with different interests their individual need for lexical simplification may be similar in some contexts, but very different when it comes to texts within their own field of interest.

Synonym pair must always be represented explicitly in the dictionary and this limits the impact of overlapping synonym sets, but does not protect against all errors. There is, for example, at present no word sense disambiguation in SynLex, and as a result the synonym sets in this dictionary will sometimes overlap.

It is very common that words have more than one sense and the particular interpretation is often realized only by the context in which the word appears. For example, the English word *deck* can be a nautical term (floor), a tire/wheel, or a pack of cards. All these possible interpretations of the word are nouns that belong to different synonym sets, and the correct interpretation of the word only becomes apparent through the context. In order to overcome this obstacle the system must be able to distinguish between word senses.

There are some electronic synonym resources for Swedish that distinguish between word senses, but the SynLex dictionary is unique in that it represents the level of synonymy of individual synonym pairs. It would be interesting to combine the SynLex dictionary with another source to create a dictionary that both respects word senses, and retains the estimated level of synonymy between the words.

The quality of the dictionary frequencies used in the study is dependent on the Swedish Parole list containing the 100.000 most common Swedish words. For the SynLex synonym dictionary the Parole frequency list was not exhaustive. A reason for this is that some of the words in SynLex are rare and simply do not appear among the most frequently used words, but another reason is that the Granska Tagger can produce erroneous lemmas, resulting in non-words which are naturally not represented in the Parole list.

If word sense disambiguation is to be supported there is also the problem of distinguishing between the frequencies of the different word senses. The Parole corpus clusters all occurrences of letter sequences only, this means the frequency counts may not be representative for the different uses of the word. For example, the Swedish verb *höra* can either mean *to*

hear or *to question someone*, but the former interpretation is much more common than the latter.

7.1.2 The inflection handler

The inflection handler is by no means state of the art and could be improved greatly. At present it cannot handle any word forms not explicitly represented in its internal dictionary of word inflections, which was generated by the Granska Tagger. It is possible that a more sophisticated morphological handler could decrease the error ratio of the synonym replacement modules, and it would likely increase the number of possible replacements. Since only synonymous alternatives that could be inflected in the same way as the original word were accepted as possible replacements the best alternatives may in some contexts have been excluded falsely due to missing inflection forms.

7.2 Implications of the experiments

The results of experiments performed in this study illustrate that viewing automatic lexical simplification as a simple task, namely that of replacing words with more common synonyms, is not in itself a sufficient strategy. There are a few studies that have dealt with lexical simplification in interesting ways. Kandula et al. (2010) estimated the familiarity of words to decide whether a word had to be simplified at all. If a certain familiarity threshold was not reached the word was simplified. If a synonym could be found that had a sufficiently high familiarity score the original word was replaced, and if no such word existed the system instead added a short explanation to the word. Using the concept of familiarity threshold could improve the quality of the texts produced by the modules in this study, since simplifying words that are already simple risks introducing errors without improving the readability of the text. Also, since rare word are more likely have a specific meaning the risk of overlapping sets of synonyms would be lower, making it less likely to introduce errors due multiple word senses.

If more information about the relationship between two synonyms was known, such as if one is a hyponym of the other, the type of short explanations generated by Kandula et al. (2010) could also be mimicked, but it could also be possible to limit other types of errors. When replacing a word with a synonym that has a more general meaning information may

be lost, but if the opposite occurs false information may be introduced to the text. If the phrase *the truck was fully loaded* was replaced by either a) *the vehicle was fully loaded* or b) *the 18-wheeler was fully loaded*, the alternative words would have different effects on the original sentence. In a) the alternative word is less specific than the original, since *truck* is a hyponym of *vehicle*, and it therefore introduces no errors, while b) specifies that it is a big truck, that is it adds information to the original expression. Avoiding the type of over-specification in b) could improve the performance of the synonym replacement.

Taking into account the surrounding words by ensuring that alternative words form standard collocations is another way of improving the quality of the generated texts. In some cases replacement words are not strictly wrong but result in sentences, or expressions, that are very uncharacteristic for native speakers of the language. But collocations van be quite complicated as the order and number of words in a collocation, and the number of intermediate words, can vary. There are some lists of collocations available, but it is not trivial to utilize this information fully. One way of dealing with such relationships is to use n-gram models or other probabilistic models. These could be used to estimate how likely a word is to appear in a particular word context.

Internet search queries is another method which could be used to test whether a replacement is likely to be correct. If a word on its own returns no hits in an Internet search query it is unlikely that the word is a real word. Furthermore, if multiple words were to be provided in a query the number of hits could be used as an estimate of the likelihood that the words form a collocation. If an alternative word forms no collocation with the surrounding words another synonym may be a better alternative, and if no collocations are found it may be better not to replace the word at all.

Chapter 8

Conclusion

Lexical simplification is a topic that requires more attention in research on automatic text simplification. A common assumption is that frequency alone is a sufficient criterion for estimating the difficulty of words. Although this is naturally not always the case word frequencies are usually a good estimate of word familiarity and, by extension, can work as an estimate of word difficulty. Based on this assumption it is easy to compare the difficulty of two words by simply referring to word frequency information. Many researchers apply this reasoning to lexical simplification but do not give appropriate attention to many of the related questions. For example, is comparison of frequency alone enough to motivate a replacement of one term for another? What if two synonymous have roughly the same frequency; will replacing one with the other affect the overall readability of a text? Also, a substitution always risks introducing an error to the text, and for automatic lexical simplification to be viable the benefit of the substitutions must outweigh the risks of introducing errors. It is therefore of importance to speak of thresholds, which allow replacements to made only if the effect of the replacement is positive. Research in this area, however, is lacking.

Within the field of natural language processing there has always been an interest in identifying the characteristics that represent text readability. Several attempts have been made to model the qualities and features necessary for describing readability in a numerical fashion. Readability metrics based on surface structure alone are the most common modelling techniques for estimating text difficulty. The calculated metrics can then be used to assess the accessibility of a text for different reader groups.

For example, the Flesh reading scale classifies a text by mapping it to a specific US grade level, and for Swedish the readability metric LIX produces a numerical value, which can be compared to a table of intervals to categorize the text in terms of the genre to which the type of language corresponds. Another popular metric is OVIX, which attempts to measure, primarily, the vocabulary load of a text. None of the established metrics take into account the actual difficulty of the terminology used in the text, though intuitively we know that a text written using short words and sentences can still be very difficult to understand.

This study indicates that the overall error ratio of replacing synonyms based on word frequencies is not affected in any significant way by the introduction of relative word frequency thresholds. But as the frequency threshold is increased we should be able to say with greater confidence that the replacements performed contribute to the overall readability of the text since the introduced words should become more and more familiar. In order to test this hypothesis one should let readers assess the readability of texts where replacements have been made with or without thresholds. It would of course be preferable if there was a readability metric that took word difficulty, or word familiarity, into account when assessing readability, but as mentioned earlier none of the established Swedish metrics do.

This study used a modified version of the SynLex dictionary to estimate the level of synonymy between word pairs. The maximum level of synonymy did not guarantee that substitutions would be correct, rather the precision at this level was still relatively poor. This suggests that the surrounding context of a word must be taken into account in order to reduce the ratio of errors for the replacements.

LIX is affected by the lengths of words in a document and studies have verified that texts with higher readability tend to use shorter words. If replacing synonyms based on word length has a positive effect on LIX this could in part be explained by the observation that word length tend to become shorter with frequency of use. If this is true modern readability metrics should try to take advantage of this. It is for example possible that word length could be used as an alternative to word frequencies when they are unknown. Length could thus potentially be used to indirectly measure how frequent a word is in a language, and by extension work as an estimate of reader familiarity.

Naturally, word length should also be considered in its own right to be a measure of text readability, since long words are more difficult to spell out for poor readers, especially if the word is not very familiar. In

some cases a combination of word frequency and word length is probably preferable. For example, *självklart* (obviously) would be replaced by *naturligtvis* (naturally) if based on word frequency, but another synonym is almost as common but also shorter *förstås* (of course). In this example the simplification could probably benefit if there was some balance between word frequency and word length. This balance could possibly be attained by establishing familiarity thresholds at which point a word is assumed to be simple enough, and after which other factors such as length or level of synonymy should determine the appropriate alternative word. These threshold could also be adapted to different reader groups.

In the study it was shown that the error ratio does not critically depend on level of synonymy. The overall error ratio remained roughly the same even at the maximum thresholds. The typical errors seem to occur because there is no disambiguation of word meaning, or because the alternative word does not fit into the particular context. Below are two examples errors introduced by the replacements from the experiments (words used as replacements are italicized):

Personer drabbade av hjärtinfarkt (People who have suffered a heart attack)
Personer *hände* av hjärtinfarkt (People who have *happened* a heart attack)
Det är lättare att gå ner i vikt (It's easier to lose weight)
Det är lättare att gå ner i *betydelse* (It's easier to lose *importance*)

The simplification system would have benefited from taking advantage of the surrounding context of each word. One method of improving the quality and precision of the modified texts would be by using some type of probabilistic model, or simple Internet search queries. A Google search query of the examples above returns a total of 6,650 hits for the phrase "drabbade av hjärtinfarkt" while the query "hände av hjärtinfarkt" returns no hits. Queries for the second example returns a total of 1,640,000 hits for "gå ner i vikt" while "gå ner i betydelse" returns only one hit. In some contexts, such as the ones above the results of simple queries clearly indicate where errors have been introduced. In order to use this type of queries in the wider sense we must however be able to choose the necessary context around a word, something which is not always easy.

Some errors are more difficult to detect using probabilistic models and search engine queries. One type of error that may be introduced involves over-specification, that is, when a replacement introduces information into the text. In order to handle this type of error it is necessary to have more fine grained information about the semantic relationship between

the original and alternative word. If the alternative word is a hyponym of the original the substitution would add information to the text, since it would be more specific, which may result in an error, but if the original word instead is a hyponym of the alternative a substitution would only make the description less specific and would be less likely to introduce actual errors. In terms of readability the reverse effect, underspecification, could also be a problem since information may be lost and the text can become more abstract, and as with synonym relations hyponym relations are often dependent upon context and the meaning of one word seldom falls completely within another.

A lot of attention should also be directed towards filtering to avoid replacing words that are already sufficiently simple, since every replacement risks introducing errors. It would be very interesting to see how sensitive readers are to the typical errors that appear as a result of the type of automatic lexical simplifications performed in this study. If the damage caused by the typical errors, in terms of readability and reading comprehension, is negligible then the precision of the replacements may be less important than the overall simplification of the substitutions. If, on the other hand, readers are sensitive to errors replacements should only be performed if words are difficult and the likelihood of a correct replacement is high.

A simplified text that contains some errors but which fails to appreciate subtle differences in terminology could still be very useful if the original text is too difficult to comprehend to the unassisted reader.

Appendix A

Manual for error evaluation

In the evaluation of modified texts two types of errors are distinguished between; type A errors, and type B errors.

Type A errors are replacements which change the semantic meaning of the sentence, introduce non-words, introduce co-reference errors within the sentence, or introduce words of the wrong class (e.g. replacement of a noun with an adjective). Replacing domain specific terminology is allowed if the alternative word's meaning corresponds closely to that of the original word (this is motivated by the fact that terminology is often more relevant for experts within a field than for non-experts). Replacements of words with expressions regarded as slang are accepted if the expressions are used correctly. This means that "isbjörnarna *softar* i solen" is an acceptable form of "isbjörnarna vilar i solen" (the polar bears are relaxing in the sun), while "*softa* vajrar gjorde att bron svajade" is not an acceptable form of "slappa vajrar gjorde att bron svajade" (loose wires made the bridge sway).

Type B errors consist of misspelled words, definite/indefinite article or modifier errors, and erroneously inflected words. The typical type B error is a word that is a correct replacement that has been written incorrectly.

Errors are either of type A or type B, and a replacement can only introduce a maximum of one error. If a replacement is of both error types it should be regarded a type A error. Also, if an error exists within the original sentence it should not be marked down as an error in the modified

sentence, since the error was not introduced by the actual replacement. For example, in some of the texts used in this study headings appear in the flow of the texts, which may produce grammatically incorrect sentences.

In some cases it may be difficult to determine whether a word should be regarded an error, therefore it is important that all modified versions of a text are assessed using the same criteria, and with the original text as a reference. If a replacement has been accepted as correct alternative in one sentence in a text it must also be accepted in all modified texts that have used the same replacement word in that sentence. If it is difficult to determine whether a replacement has introduced an error the rater is recommended write a comment about the sentence for future reference. Functionality for commenting replacements is not implemented in the program for creating answer sheets, instead comments are written in a separate document. The examples below demonstrate common types of errors:

Example 1

Anders gick över vägen för att hämta en liten sten.
Anders *promenera genom* vägen för att hämta en liten *gruskorn*.

The modified sentence contains one error of type A (the preposition *genom*), and two errors of type B (wrong tense of *promenera/promenerade*, and article errors introduced by *gruskorn*, where *en liten* should have been *ett litet*, which counts only as one error).

Example 2

Den vita musen rusade over bordsskivan.
Den *ljusa gnagaren sprang* över *bordet*.

The modified sentence does not contain any obvious errors in itself, since the general meaning of the sentence is maintained. If the rater, with the use of the surrounding context, finds that one or more of the replaced words are errors a motivation is required.

Bibliography

Aaron, P G; Joshi, M, and Williams, K A. Not all reading disabilities are alike. *Journal of Learning Disabilities*, 32(2):120–137, 1999. URL http://ldx.sagepub.com/cgi/doi/10.1177/002221949903200203.

Blake, Catherine; Kampov, Julia; Orphanides, Andreas K; West, David, and Lown, Cory. Unc-ch at duc 2007: Query expansion, lexical simplification and sentence selection strategies for multi-document summarization. *Proceedings of Document Understanding Conference (DUC) Workshop 2007*, 2007.

Bolshakov, Igor A. and Gelbukh, Alexander. Synonymous paraphrasing using wordnet and internet. *Natural Language Processing and Information Systems*, pages 189–200, 2004. URL http://www.springerlink.com/index/57LBVKHLYMJQAGJ3.pdf.

Borin, Lars and Forsberg, Marcus. All in the family: A comparison of saldo and wordnet. 2009. URL http://hdl.handle.net/10062/9836.

Carroll, John; Minnen, Guido; Canning, Yvonne; Devlin, Siobhan, and Tait, John. Practical simplification of english newspaper text to assist aphasic readers. *Proceedings of the AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1(11):7–10, 1998. URL http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract.

Carroll, John; Minnen, Guido; Pearce, Darren; Canning, Yvonne; Devlin, Siobhan, and Tait, John. Simplifying text for language-impaired readers. In *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 269–270, 1999.

Chandrasekar, R. and Srinivas, B. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190, 1997. ISSN 0950-7051. doi: 10.1016/S0950-7051(97)00029-4. URL http://www.sciencedirect.com/science/article/pii/S0950705197000294.

Chandrasekar, R.; Doran, Christine, and Srinivas, B. Motivations and methods for text simplification. In *PROCEEDINGS OF THE SIXTEENTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING '96)*, 1996.

Dana, Dannélls. Automatic generation and simplification of written documents. Accepted to the Sixth International and Interdisciplinary Conference on Modeling and Using Context, Roskilde University, Denmark, 2007. URL http://spraakdata.gu.se/svedd/pub/context07.pdf.

Decker, Anna. Towards automatic grammatical simplification of swedish text. Master's thesis, Stockholm's University, Sweden, 2003.

Dell'Orletta, Felice; Montemagni, Simonetta, and Venturi, Giulia. Read-it: assessing readability of italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pages 73–83, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-14-5. URL http://dl.acm.org/citation.cfm?id=2140499.2140511.

Domeij, Rickard; Knutsson, Ola; Carlberger, Johan, and Kann, Viggo. Granska - an efficient hybrid system for swedish grammar checking. In *Proceedings of the 12th Nordic Conference in Computational Linguistics, Nodalida-99*, 2000.

Feng, Lijun; Elhadad, Noémie, and Huenerfauth, Matt. *Cognitively motivated features for readability assessment*, pages 229–237. Number April. Association for Computational Linguistics, 2009. URL http://portal.acm.org/citation.cfm?doid=1609067.1609092.

Kandula, Sasikiran; Curtis, Dorothy, and Zeng-Treitler, Qing. A semantic and syntactic text simplification tool for health content. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, pages 366–370, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041424&tool=pmcentrez&rendertype=abstract.

Kann, Viggo and Rosell, Magnus. Free construction of a free swedish-dictionary of synonyms. In *NoDaLiDa 2005*, pages 1–6, 2005. QC 20100806.

Kokkinakis, Dimitrios; Gronostaj, Maria Toporowska, and Johansson Kokkinakis, Sofie. Att bygga en sprakbro mellan allmanhet och vardpersonal - spraket i texter om hjart-karlsjukdomar. Goteborg's University, Sweden, 2006.

Köster-Bergman, Lena. Sverige läser bäst i världen - men..., 2001. URL `http://www.fungerandemedier.se/sites/fungerandemedier.se/files/pdf/Sverige%20l%C3%A4ser%20b%C3%A4st%20i%20v%C3%A4rlden.pdf`.

Lal, Patha and Rüger, Stefan. *Extract-based summarization with simplification*. 2002. URL `http://www.mariapinto.es/ciberabstracts/Articulos/extract-based.pdf`.

Luhn, H P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(April):159–165, 1958.

Miller, George A. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

Mühlenbock, Katarina and Johansson Kokkinakis, Sofie. Lix 68 revisited an extended readability measure. *Focus*, pages 1–9, 2010.

Rybing, Jonas; Smith, Christian, and Silvervarg, Annika. Towards a rule based system for automatic simplification of texts. Swedish Language Technology Conference, SLTC, Linköping, Sweden, 2010. URL `http://www.ida.liu.se/~chrsm81/papers/sltc10.pdf`.

Siddharthan, Advaith. Syntactic simplification and text cohesion. Technical report, Research on Language and Computation, 2003.

Siddharthan, Advaith and Copestake, Ann. *Generating Anaphora for Simplifying Text*, pages 199–204. Number Daarc. 2002.

Smith, Christian and Jönsson, Arne. Automatic summarization as means of simplifying texts, an evaluation for swedish. *Evaluation*, 2007. URL `http://www.ida.liu.se/~arnjo/papers/nodalida-11.pdf`.

Wei, Xing; Peng, Fuchun; Tseng, Huihsin; Lu, Yumao, and Dumoulin, Benoit. Context sensitive synonym discovery for web search queries. *Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09*, page 1585, 2009. URL `http://portal.acm.org/citation.cfm?doid=1645953.1646178`.

LINKÖPINGS UNIVERSITET

Avdelning, Institution
Division, Department

IDA,
Dept. of Computer and Information Science
581 83 LINKÖPING

Datum
Date

2012-10-09

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English
☐ _____

Rapporttyp
Report category

☐ Licentiatavhandling
☒ Examensarbete
☐ C-uppsats
☐ D-uppsats
☐ Övrig rapport
☐ _____

ISBN
—

ISRN
LIU-IDA/KOGVET-A–12/014–SE

Serietitel och serienummer **ISSN**
Title of series, numbering —

URL för elektronisk version
http://www.ep.liu.se/exjobb/ida/202012/dd-d/LIU-IDA/KOGVET-A--12/014--SE/

Titel Automatiskt textförenkling genom synonymutbyte

Title Automatic Text Simplification via Synonym Replacement

Författare Robin Keskisärkkä
Author

Sammanfattning
Abstract

In this study automatic lexical simplification via synonym replacement in Swedish was investigated using three different strategies for choosing alternative synonyms: based on word frequency, based on word length, and based on level of synonymy. These strategies were evaluated in terms of standardized readability metrics for Swedish, average word length, proportion of long words, and in relation to the ratio of errors (type A) and number of replacements. The effect of replacements on different genres of texts was also examined. The results show that replacement based on word frequency and word length can improve readability in terms of established metrics for Swedish texts for all genres but that the risk of introducing errors is high. Attempts were made at identifying criteria thresholds that would decrease the ratio of errors but no general thresholds could be identified. In a final experiment word frequency and level of synonymy were combined using predefined thresholds. When more than one word passed the thresholds word frequency or level of synonymy was prioritized. The strategy was significantly better than word frequency alone when looking at all texts and prioritizing level of synonymy. Both prioritizing frequency and level of synonymy were significantly better for the newspaper texts. The results indicate that synonym replacement on a one-to-one word level is very likely to produce errors. Automatic lexical simplification should therefore not be regarded a trivial task, which is too often the case in research literature. In order to evaluate the true quality of the texts it would be valuable to take into account the specific reader. A simplified text that contains some errors but which fails to appreciate subtle differences in terminology can still be very useful if the original text is too difficult to comprehend to the unassisted reader.

Nyckelord
Keywords Lexical simplification, synonym replacement, SynLex

Copyright

Svenska

English