Northern European Journal of Language Technology, 2013, Vol. 3, Article 3, pp 41–59
DOI 10.3384/nejlt.2000-1533.1333

Investigations of Synonym Replacement for Swedish

Robin Keskisärkkä, Arne Jönsson
Santa Anna IT Research Institute AB
Linköping, Sweden
robin.keskisarkka@liu.se, arnjo@ida.liu.se

Abstract

We present results from an investigation on automatic synonym replacement for Swedish. Three different methods for choosing alternative synonyms were evaluated: (1) based on word frequency, (2) based on word length, and (3) based on level of synonymy. These three 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 in relation to replacements. The results show an improvement in readability for most strategies, but also show that erroneous substitutions are frequent.

1 Introduction

Synonym replacement is a technique that can make a text easier to read. When replacements are based on word length the text will contain a lower number of long words, and if words are replaced by simpler synonyms there will be a smaller variation in terms of unique words, since multiple nuanced words may be replaced by the same word. In both cases readability would be improved in terms of established readability metrics. But metrics alone do not tell the whole story about the readability or quality of a text.

Familiarity and the perceived difficulty of a word is related to how often an individual is exposed to it. In the Swedish Parole list of frequencies the words *icke* and *inte* (both meaning "not") differ considerably in how frequently they are used; *icke* has a frequency of 1,244, and *inte* has a frequency of 183,952. This accurately reflects that the former is more old fashioned, and is normally considered to be more bureaucratic. But the difference in frequency between two words is often less notable, as in the case of *allmän* (public), with a frequency of 686, and its possible synonym *offentlig* (official), with a frequency of 604. Does the relatively small difference in frequency mean that one is easier to understand than the other? Words can also, 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).

Words with identical meaning are rare and any tool that replaces words automatically is therefore likely to affect the content of the text. This, however, does not mean that automatic lexical simplification could not be useful, e.g., persons with limited knowledge of economy may profit little from the distinction between the terms *income*, *salary*, *profit*,

and *revenue*. Replacing these terms with a single word, say *income*, would result in a document that fails to appreciate the subtle differences between these concepts, but it does not necessarily affect an individual's understanding of the text to the same degree, if the words appear in context.

The aim of this study can be summarized into two main questions: 1) To what degree can automatic lexical simplification on the level of one-to-one synonym replacement be successfully applied to Swedish texts? and 2) Can thresholds for replacements be introduced to maximize the quality of the simplified texts?

2 Background

In our work we will do lexical simplifications based on synonymy between single words. The degree of success will be measured based on readability and, as presented in Section 3, on the number of erroneous replacements.

2.1 Lexical simplification

Lexical simplification of written text can be accomplished using various strategies. 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.

A way of performing lexical simplification was implemented by Carroll et al. (1998, 1999) in a simplifier that used word frequency counts to estimate the difficulty of words. Their system passed words one at 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 more 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 uses a modular architecture where components can 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 witDh low familiarity scores, identified by a combination of the words' usage contexts and its frequency in layperson reader targeted biomedical sources. The familiarity score as an estimate of word difficulty was successfully validated using customer surveys. Their definition of familiarity score resulted 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 instead added to the text. A simple explanation phrase was generated based on the relationship between the difficult term and a related term with higher familiarity score. An explanation was either of the form <difficult_term> (a type of <parent>), or of the form <difficult_term> (e.g. <child>), depending on the relationship between the two words. An earlier study had showed that these two types of relations produced useful and correct explanations in 68% of the generated explanations. The authors also introduced additional non-hierarchical semantic explanation connectors in their study.

Another lexical simplification technique is to remove sections of a sentence that convey non-essential information, or superfluous words. This technique has, e.g., been used to simplify texts to improve automatic text summarization (Blake et al., 2007).

2.2 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, e.g., *United States of America*, *United States*, *USA*, and *US*. Absolute synonyms can occur in the same context without affecting the overall style or meaning of the text, but such equivalence relations are extremely rare in all languages. Bolshakov and Gelbukh suggested that the inclusion of multiword and compound expressions in synonym databases, however, would result in 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 is considered synonymous with the other words in that particular set (Miller, 1995). This builds on the assumption that synonymy is a symmetric property, that is, if *car* is synonymous with *vehicle* then *vehicle* should also be regarded synonymous with *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 a synonym of $word_3$ then $word_1$ and $word_3$ can be viewed as synonyms (Siddharthan and Copestake, 2002). This view was not used in this study, since overlapping groups of synonyms can result in extremely large synsets, especially if word sense disambiguation is not applied. The view of synonymy as a symmetric and transitive property is seldom discussed in the literature, but it is closely related to the distinction of hyponyms.

Hyponyms express a hierarchical relation between two semantically related words, e.g., in the previous example *car* can be viewed as 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 like absolute synonyms true hyponym relations are rare. The two words above can therefore sometimes be viewed as synonymous, but in most cases *vehicle* has a more general meaning. Replacement of the term *car* for *vehicle* would thus, in most contexts, produce a less precise description but would likely not introduce any errors. However, if the opposite were to occur, that is, if *vehicle* would be replaced by *car*, the description

would become more specific and would run a higher risk of producing errors. In practise, many words can not 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 a 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.

2.3 Readability metrics

This section briefly defines the established readability metrics for Swedish and the textual properties that they tend to reflect. Theoretically, synonym replacements can affect established readability metrics different ways. The correlation between word length and text difficulty indicates that lexical simplification is likely to result in decreased word length overall, and a decrease in number of long words. Also, if words are replaced with simpler synonyms we can expect fewer unique words, since multiple nuanced words may be replaced by the same word.

For Swedish, being an inflecting and compounding language, the readability index LIX (Björnsson, 1968) is the most frequently used readability measure. Other common measures are, OVIX, AWL and LWP (Mühlenbock and Kokkinakis, 2009).

LIX combines the average number of words per sentence, and the proportion of long words in the text, Equation 1. A high number indicates a more complicated text. Lexical variation, or OVIX (word variation index), measures the ratio of unique tokens, Equation 2. The OVIX-value functions as a metric of vocabulary load. Average word length, AWL, is calculated by dividing the number of characters in a text by the number of words, Equation 3. Finally, the ratio of long words, LWP, is the number of words with more than six characters divided by all words, Equation 4. In the equations below $n(x)$ denotes the number of x.

$$LIX = \frac{n(words)}{n(sentences)} + (\frac{n(words\ >\ 6\,chars)}{n(words)} \times 100) \tag{1}$$

$$OVIX = \frac{log(n(words))}{log(2 - \frac{log(n(unique\ words))}{log(n(words))})} \tag{2}$$

$$AWL = \frac{n(characters)}{n(words)} \tag{3}$$

$$LWP = \frac{n(words\ >\ 6\,chars)}{n(words)} \tag{4}$$

3 Method

In this section we present the software modules developed for the project and the language resources they use. We also present the experimental procedure.

3.1 Language resources

We use the freely available SynLex in which level of synonymy between words is represented in the interval 3.0–5.0, where higher values indicate a higher degree of synonymy between words. The lexicon was constructed by having Internet users of the Lexin translation service rate the level of synonymy between Swedish words on a scale from one to five (Kann and Rosell, 2005). Users of the service 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 entered into the research set. The average levels of synonymy were summarized when 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.

SynLex was combined with Parole's frequency list of the 100,000 most common Swedish words. The Granska Tagger (Domeij et al., 2000), a part-of-speech tagger for Swedish, was used to generate the lemma forms of each of the words in the frequency list, and the frequency counts for each identical lemma collapsed into a more representative list of word frequencies. The lemma frequencies for words based on this list were added as an attribute to each word in the list of synonyms. If a word was not in the frequency list it was listed as zero and the word pair was excluded from the synonym list. The final look-up table contained synonym pairs in lemma form, level of synonymy between word pairs, and word frequency count for each word. The original synonym list contained a total of 37,969 synonym pairs, but after adding frequency, and excluding words with a word frequency of zero, 23,836 pairs remained.

Fewer synonym pairs may have been lost if the entire Parole frequency list had been used, rather than limiting it to the 100,000 most common words, which included only those words with frequency counts of 6 or more. However, since some of the entries in SynLex were multiword expressions, and frequencies were available only for unigrams, some of the synonyms in SynLex would still be listed with a frequency count of zero.

3.2 System

Three main modules were developed in Java. In the first module, replacements were performed based on word frequency counts, as an estimate of word familiarity. In the second module, replacements were performed based on word length, motivated by established readability metrics, which state that word length correlates with readability of a text. In the third module, words were replaced with the synonym having the highest level of synonymy.

An inflection handler was also developed, which enabled word forms of lemmas to be looked up quickly by using lemma and inflection information as parameters, which could be generated for a word using the Granska Tagger. The modules were modified to generate lemma and word class information for each word in a text, and to look for

synonyms based on this information. If the original word's inflection form could not be generated for the alternative word it was not allowed as a replacement.

Finally, the option to add thresholds were introduced for each criterion in the modules; if the criterion threshold value was not reached the replacement would not take place.

The techniques employed by the modules can produce a variety of errors, e.g., deviations from the original semantic meaning, replacement of established terminology, formation of strange collocations, deviation from general style, and syntactic or grammatical incorrectness. For the purpose of this study the errors were clustered into two separate categories: *Type A errors* include replacements which change the semantic meaning of the sentence, introduces non-words, introduces co-reference errors within the sentence, or introduces a different word class (e.g. replaces a noun with an adjective). *Type B errors* consist of misspelled words, article or modifier errors, and erroneously inflected words. Errors not included in the two error types, e.g., stylistic errors, were ignored to minimize the effects of subjectivity in the rating of texts, minimize the effects of a rater's domain knowledge, and to simplify the rating procedure. The texts were checked for errors manually, using a predefined manual describing errors constituting Type A and Type B errors. The inter-rater reliability between two raters using the manual was 91.3%.

3.3 Procedure

Sixteen texts were chosen from four different genres: newspaper articles from *Dagens nyheter* (DN), informative texts from the Swedish Social Insurance Administration's (Försäkringskassan) homepage (FOKASS), popular science articles from *Forskning och framsteg* (FOF), and academic text excerpts (ACADEMIC). Every genre consisted of four different documents, which were of roughly the same size. The average text contained 54 sentences and each sentence contained on average 19 words. In the experiments, synonym replacement was performed on the texts using a one-to-one matching between all words in the original text and the available synonyms. A filter was used 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.

In the first two experiments the conditions word frequency, word length, and level of synonymy are used to choose the best replacement alternatives. The first compares word frequencies and performs substitutions only if the alternative word's frequency is higher than that of the original, if more than one word meets this criterion the one with the highest word frequency is chosen. The second replaces a word only if the alternative word is shorter, and if more than one word meets the criterion the shortest one is chosen. The third replaces every word with the synonym that has the highest level of synonymy. In experiment 2 the inflection handler is introduced. The inflection handler allows synonym replacement to be performed based on lemmas, which increases the number of potential replacements. The inflection handler also functions as an extra filter for the replacements, since only words that have an inflection form corresponding to that of the word being replaced are considered as alternatives. In the third experiment thresholds are introduced for the different strategies. The thresholds are increased incrementally and errors are evaluated for each new threshold. Finally, in the fourth experiment word frequency and level of synonymy are combined and used with predefined thresholds.

4 Results

This section presents the results of the experiments performed in this study.

4.1 Experiment 1: Synonym replacement

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 no inflection handler was included only words written in their lemma form were evaluated for substitution.

4.1.1 Synonym replacement based on word frequency

The results presented in Table 1 show that the replacement strategy based on word frequency resulted in a significant improvement in almost all readability metrics for every genre, and for the texts in general.

Table 1: Average LIX, OVIX, LWP, and 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 are presented in Table 2. The results show that the amount of erroneous replacements is very high, on average more than half of all replacements have been marked as errors, 0.52. 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 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)

4.1.2 Synonym replacement based on word length

The results presented in Table 3 show that the replacement strategy based on word length resulted in a significant improvement in terms of readability for every genre, and for the texts in general, in almost all readability metrics.

Table 3: Average LIX, OVIX, LWP, and AWL for synonym replacement based on word length. Numbers as in Table 1.

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 4. The results show that the amount of erroneous replacements for this module is very high. 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 the 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 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)

4.1.3 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 to understand than the original, however, the results are relevant as a reference in the discussion to follow as it can be seen as a test of the precision of the synonym dictionary. The results in Table 5 show that for all genres the replacement based on level of synonymy affected the readability metrics negatively except for the OVIX-value.

The errors produced by the module are presented in Table 6. The results show that the amount of erroneous replacements is high. A one-way ANOVA was used to test for differences among the 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: Average LIX, OVIX, LWP, and AWL for synonym replacement based on level of synonymy. Numbers as in Table 1.

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)

Table 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)

4.2 Experiment 2: Synonym replacement with inflection handler

In experiment 2 the inflection handler was introduced. Its function is 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.

4.2.1 Synonym replacement based on word frequency

The results presented in Table 7 show that the replacement strategy based on word frequency resulted in an improvement in terms of all readability for the texts in general in all readability metrics, and significant for some metrics when looking at the individual genres.

Table 7: Average LIX, OVIX, LWP, and AWL for synonym replacement based on word frequencies with inflection handler. Numbers as in Table 1.

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 are presented in Table 8. The results show that the amount of erroneous replacements is high. A one-way ANOVA was used to test for

differences among the 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 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)

4.2.2 Synonym replacement based on word length

The results presented in Table 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.

Table 9: Average LIX, OVIX, LWP, and AWL for synonym replacement based on word length with inflection handler. Numbers as in Table 1.

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 are presented in Table 10. The results show that the amount of erroneous replacements for this module is high. A one-way ANOVA was used to test for differences among the 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 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)

4.2.3 Synonym replacement based on level of synonymy

The results in Table 11 show that for all genres the replacement based on level of synonymy affected the readability metrics negatively for all genres and all metrics except for the OVIX-value for ACADEMIC.

Table 11: Average LIX, OVIX, LWP, and AWL for synonym replacement based on level of synonymy with inflection handler. Numbers as in Table 1.

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 are presented in Table 12. The results show that the amount of erroneous replacements is high. A one-way ANOVA was used to test for differences among the 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 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)

4.3 Experiment 3: Threshold estimation

In experiment 3 thresholds were introduced and incrementally increased to see if there were any relationships between the level at which a synonym was accepted as a replacement word and the error ratio. All replacements run the risk of introducing a Type A error, therefore the benefit of a replacement should be viewed relative to the effect it has on the readability of the text. By introducing thresholds for a replacement, effects of change in a criterion can be studied. For further details see Keskisärkkä (2012).

4.3.1 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 increasing 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.

We found no clear relationship between threshold and error ratio when all texts were analyzed together. For some texts the error ratio decreases as the threshold increases, but for others the opposite is true. Clustering particularly occurs near the maximum values of the two variables.

For the texts in their respective genres a weak, but significant, correlation between threshold and error ratio was found for ACADEMIC, $r(234) = -.205, p < .01$, DN, $r(234) = -.231, p < .001$, and a positive correlation was found for FOF, $r(234) = .197, p < .01$. As before, the result of this experiment depends almost exclusively on the nature of individual texts, rather than on which genre it belongs to.

4.3.2 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, removing first those replacements which only result in slightly shorter words.

We found no clear relationship between threshold and error ratio when all texts are analyzed together. The threshold and error ratio summarized for the texts in their respective genres shows 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 texts, rather than on which genre it belongs to, only in FOKASS the relationship was strong.

4.3.3 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 level of synonymy.

We found no clear relationship between threshold and error ratio when all texts were analyzed together. The threshold and error ratio summarized for the texts in their respective genres show that a significant negative correlation between threshold and error ratio exists for DN, $r(82) = -.498, p < .001$, and that 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 texts, rather than on which genre it belongs to.

4.4 Experiment 4: Frequency combined with level of synonymy thresholds

Experiment 3 revealed no clear relationship between threshold and error ratio for any of the three replacements strategies. For some texts the error ratio decreased as the threshold increased, while for others the opposite was true, and the ratio of errors remained

relatively constant. However, this does not necessarily mean that the thresholds do not have any affect. There is still room for interaction effects when two or more strategies are combined. 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 analysis of errors must be performed. Instead, the word frequency strategy, which had the best error ratio in experiment 2 and which is the strategy with the strongest support in the research literature, was combined with level of synonymy. The change in frequency threshold did not change the average error ratio, however, the motivation for synonym replacement based on word frequency is that the alternative word should be significantly more familiar than the original word. For this experiment the frequency threshold was set to 2.0, meaning that only replacements with a frequency count at least two times that of the original word were accepted as alternatives. At the same time the quality of the synonym must be high. For this experiment the threshold for the minimum level of synonymy was set to 4.0.

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 this experiment both of these strategies were investigated.

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 two. The experiment was run twice, prioritizing either frequency (PrioFreq) or level of synonymy (PrioLevel) when more than one synonym passed the thresholds. The words replaced are 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 significant differences in terms of error ratio, whether comparing all texts 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).

5 Discussion

The results from our study show that synonym replacement generally improves readability in terms of the readability measures used. This is independent of genre and if replacements are based on word frequency or word length. Replacements based on level

of synonymy affect readability negatively, but since replacements are performed regardless of whether the new word is easier to understand or not this is not surprising.

The study shows that the common view of automatic lexical simplification as a task of simply replacing words with more common synonyms results in a lot of erroneous replacements. The error ratio does not critically depend on level of synonymy, rather the overall error ratio remains roughly the same even when using words with the highest level of synonymy. The error ratio decreases if we consider inflections, but are still high. The high error ratios at this level confirm that the concept of synonyms relies greatly on the context. Handling word collocations and word sense disambiguation could greatly improve both the quality of the modified texts and substantially decrease the error ratio, but this is by no means a trivial task.

A way of handling this problem is to present potential synonyms to a user without actually replacing the words. This approach is currently integrated in an automatic summarizer (Smith and Jönsson, 2011), where the overall goal is to simplify texts in order to facilitate information access.

In the study we found no general thresholds, in any of the strategies, for which the ratio between errors and replacements improves significantly. The overall error ratio of replacing synonyms based on frequency is not significantly affected by the introduction of relative threshold frequencies, but increased threshold means that the new words are more likely to be familiar to the reader. Word length as a strategy for synonym replacement improves the text in terms of the readability metrics, but it is not clear whether it actually contributes to the readability of the text. Also, the combination of frequency and level of synonymy only slightly improves the error ratio compared to frequency alone. In summary, the likelihood that the replacement is a correct synonym in the particular context does not increase significantly with thresholds, but the likelihood that the replacement word is easier to read does.

Synonym replacement based on length, frequency and level of synonymy are applicable to many languages. Measures of readability differ between languages, but the use of LIX has not been limited to Swedish, it has also been used to assess readability in both Danish and Norwegian texts (Delsing and Lundin-Åkesson, 2005). The three languages greatly resemble each other in terms of both vocabulary and grammar, which suggests that the results of this study may be generalizable to the other Scandinavian languages.

It is known that the established readability metrics for Swedish have some shortcomings (Mühlenbock and Kokkinakis, 2009). A metric that could take into account word difficulty, e.g., how common a word is (Falkenjack and Mühlenbock, 2012), in a more direct fashion would better describe how a text is affected by a simplification via one-to-one replacements. Another alternative would be to assess the readability of texts using actual readers. A simplified text can potentially be useful even if it contains some errors, especially if the original text is too difficult to comprehend for the unassisted reader. It would therefore be interesting to study the sensitivity of readers to typical errors, and the investigate the effects this type of simplification has on reading comprehension. Future work should also aim at replacing only those words which are regarded difficult to a particular reader, rather than trying to simplify all words.

References

Björnsson, C.H. 1968. *Läsbarhet*. Stockholm: Liber.

Blake, Catherine, Julia Kampov, Andreas K Orphanides, David West, and Cory Lown. 2007. 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* .

Bolshakov, Igor A. and Alexander Gelbukh. 2004. Synonymous Paraphrasing Using WordNet and Internet. *Natural Language Processing and Information Systems* pages 189–200.

Borin, Lars and Marcus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series*.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, vol. 1, pages 7–10. Citeseer.

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

Delsing, Lars-Olof and Katarina Lundin-Åkesson. 2005. *Håller språket ihop Norden? : en forskningsrapport om ungdomars förståelse av danska, svenska och norska*. TemaNord. Nordiska ministerrådet.

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

Falkenjack, Johan and Katarina Heimann Mühlenbock. 2012. Using the probability of readability to order Swedish texts. In *Proceedings of the Fourth Swedish Language Technology Conference, Lund, Sweden*.

Kandula, Sasikiran, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A Semantic and Syntactic Text Simplification Tool for Health Content. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium* pages 366–370.

Kann, Viggo and Magnus Rosell. 2005. Free Construction of a Free Swedish Dictionary of Synonyms. In *NoDaLiDa 2005*, pages 1–6. QC 20100806.

Keskisärkkä, Robin. 2012. *Automatic Text Simplification via Synonym Replacement*. Master's thesis, Linköping University, Department of Computer and Information Science.

Lal, Patha and Stefan Rüger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL*.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38:39–41.

Mühlenbock, Katarina and Sofie Johansson Kokkinakis. 2009. LIX 68 revisited – An extended readability measure. In *Proceedings of Corpus Linguistics*.

Siddharthan, Advaith and Ann Copestake. 2002. Generating Anaphora for Simplifying Text. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium DAARC 2002*, pages 199–204.

Smith, Christian and Arne Jönsson. 2011. Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia*.

Wei, Xing, Fuchun Peng, Huihsin Tseng, Yumao Lu, and Benoit Dumoulin. 2009. Context sensitive synonym discovery for web search queries. *Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09* page 1585.