Synonym Replacement based on a Study of Basic-level Nouns in Swedish Texts of Different Complexity

Evelina Rennes, Arne Jönsson
Department of Computer and Information Science
Linköping University, Linköping, Sweden
evalina.rennes@liu.se, arne.jonsson@liu.se

Abstract

In this article, we explore the use of basic-level nouns in texts of different complexity, and hypothesise that hypernyms with characteristics of basic-level words could be useful for the task of lexical simplification. Basic-level terms have been described as the most important to human categorisation. They are the earliest emerging words in children's language acquisition, and seem to be more frequently occurring in language in general. We conducted two corpus studies using four different corpora, two corpora of standard Swedish and two corpora of simple Swedish, and explored whether corpora of simple texts contain a higher proportion of basic-level nouns than corpora of standard Swedish. Based on insights from the corpus studies, we developed a novel algorithm for choosing the best synonym by rewarding high relative frequencies and monolexemity, and restricting the climb in the word hierarchy not to suggest synonyms of a too high level of inclusiveness.

1 Introduction

The research concerned with automatically reducing the complexity of texts is called *Automatic Text Simplification* (ATS). Automatic text simplification was first proposed as a pre-processing step prior to other natural language processing tasks, such as machine translation or text summarisation. The assumption was that a simpler syntactic structure would lead to less ambiguity and, by extension, a higher quality of text processing (Chandrasekar et al., 1996). However, one of the main goals of modern automatic text simplification systems is to aid different types of target readers. The manual production of simple text is costly and if this process could be automated, this would have a beneficial effect on the targeted reader, as well as the society as a whole. Previous ATS studies have targeted different reader groups, such as second language (L2) learners (Petersen and Ostendorf, 2007; Paetzold, 2016), children (De Belder and Moens, 2010; Barlacchi and Tonelli, 2013; Hmida et al., 2018), persons with aphasia (Carroll et al., 1998; Canning and Tait, 1999; Devlin and Unthank, 2006), the hearing-impaired (Inui et al., 2003; Daelemans et al., 2004; Chung et al., 2013), and other persons with low literacy skills (Aluísio et al., 2008; Candido Jr et al., 2009; Aluisio et al., 2010). Reducing the complexity of a text can be done in numerous ways but one of the subtasks of ATS is lexical simplification: the process of finding and replacing difficult words or phrases with simpler options. Finding such simpler words can be done by using frequency measures to choose between substitution candidates with the intuition that the more common a word is, the simpler a synonym it is. As pointed out, for instance by Alfter (2021), more frequent words can also be complex as they tend to be more polysemous.

Finding simpler words can also be done by studying how human writers do. To write simple texts, the writers usually consult guidelines. For Swedish, such guidelines are given by Myndigheten för Tillgängliga Medier (MTM)[1]. The MTM guidelines state, among other things, that the text should be adapted to the type of reader who is going to read the text, and that everyday words should be used (MTM, 2020).

In this article, we explore the use of basic-level nouns in texts of different complexity, and hypothesise that hypernyms with characteristics of basic-level words could be useful for the task of lexical simplification. We then use this knowledge to cre-

[1] Swedish Agency for Accessible Media

ate an algorithm for synonym replacement. The conventional definition of a *synonym* is a word that have the same or nearly the same meaning as another word. However, for simplicity, in this article we extend this notion to also include near-synonyms or other semantically similar words.

Hypernyms have been previously studied from the perspective of lexical simplification. For example, Drndarević and Saggion (2012) explored the types of lexical simplification operations that were present in a parallel corpus comprising 200 standard and simple news texts in Spanish, and found that the exchanged words could be hypernyms, hyponyms and meronyms. Biran et al. (2011) used the vocabularies of Wikipedia and Simple English Wikipedia to create word pairs of content words, and one of the methods for filtering out substitution word pairs was to consult the synonym and hypernym relations between the words. Comparable synonym resources for Swedish include SynLex (Kann and Rosell, 2005) and Swe-Saurus (Borin and Forsberg, 2014).

Given what we know how simple texts are written, it seems probable that a corpus of simple text, targeting children and readers with different kinds of disabilities, is characterised by a higher proportion of basic-level nouns than, for example, a corpus comprising texts that are said to reflect general Swedish language of the 90's. The aim of this study was to explore this claim in corpora of simple and standard texts, and to see how this could be used in the context of lexical text simplification.

2 Basic-level Words

Prototype theory, as defined by Rosch et al. (1976), claims that there is a scale of human categorisation where some representing concepts are more representative than others. For example, *furniture* can be regarded as higher up in the taxonomy than *chair* or *table*, whereas *kitchen chair* or *dining table* can be found at a lower level with higher specificity. Rosch et al. (1976) found that the basic level is the most important to human categorisation. For example, basic-level terms emerge early in a child's language acquisition, and such terms generally seem to be more frequently occurring in language. Another characteristic of basic-level terms is that they often comprise one single lexeme, while subordinate terms more often consist of several lexemes (Evans, 2019).

Theories in cognitive linguistics are important for computational linguists as they adopt a usage-based approach. This means that language use is essential to how our knowledge of language is gained, and plays a large role in language change and language acquisition (Evans, 2019). When a child learns a language, the knowledge is gathered through extraction of constructions and patterns, a process grounded in general cognitive processes and abilities. One of the central ideas in the usage-based approach is that the relative frequency of linguistic constructions (such as words) affects the language system so that more frequent constructions are better entrenched in the system, thus further influencing language use.

Within the field of cognitive linguistics corpora is one of the proposed methods to study language (Evans, 2019). Corpora make it relatively simple to perform large-scale analyses in order to get quantitative measures on how language is used in a naturalistic setting. The simplest measures we can use are frequency counts, which can provide insights in how commonly used certain constructions are, in comparison with others.

3 Corpus Analysis

We conducted two corpus studies using different corpora.

The first study aimed to compare two corpora, where the first corpus contained texts that reflect the Swedish language, and the second corpus contained easy-to-read texts. The *Stockholm-Umeå Corpus (SUC)* corpus (Ejerhed et al., 2006) is a balanced corpus of Swedish texts written in the 1990's. In this study, we used the 3.0 version of the corpus (*SUC3*).

The *LäSBarT* corpus (Mühlenbock, 2008), is a corpus of Swedish easy-to-read texts of four genres: easy-to-read news texts, fiction, community information, and children's fiction. The *LäSBarT* corpus was compiled in order to mirror simple language use in different domains and genres but it is not truly balanced in the traditional sense.

The hypothesis was that the *SUC3* corpus would exhibit a higher average number of steps to the top-level noun than the *LäSBarT* corpus.

The second study aimed to investigate whether the genre did play a role. In order to investigate this, we conducted an analysis of a corpus of the Swedish newspaper *8 Sidor*, that comprises news articles in Simple Swedish, and a corpus with Göteborgs-Posten articles (*GP2D*). The corpora were of the same genre, but not parallel.

The hypothesis was that the *GP2D* corpus would exhibit an even higher average number of steps to the top-level noun than the *8 Sidor* corpus. The SUC3 corpus is balanced and, hence, also includes, for instance, simple texts that may affect the difference between the corpora.

3.1 Procedure

All nouns of the resources were extracted, together with their most probable sense gathered from SALDO (Svenskt Associationslexikon) version 2 (Borin et al., 2008). SALDO is a descriptive lexical resource that, among other things includes a semantic lexicon in the form of a lexical-semantic network.

SALDO was also used for extracting lexical relations. For each such noun, we recursively collected all *primary parents* of the input word. The *primary* descriptor describes an entry which better than any other entry fulfils two requirements: (1) it is a semantic neighbour of the entry to be described (meaning that there is a direct semantic relationship, such as synonymy, hyponymy, and meronymy, between words); and (2) it is more central than the given entry. However, there is no requirement that the *primary* descriptor is of the same part of speech as the entry itself.

The number of steps taken to reach the top-level noun was counted. The algorithm ended when there were no more parents tagged as a noun. The method was inspired by the collection of synonym/near-synonym/hypernym relations in Borin and Forsberg (2014).

In addition to this analysis, we also collected the frequency counts of the nouns occurring in the corpora and their superordinate nouns, as well as indication of compositionality. The frequency measures used were relative frequencies gathered from the *WIKIPEDIA-SV* corpus, accessed through Språkbanken[2].

3.2 Corpus Analysis Results

The number of extracted instances were 206,609 (*SUC3*), 177,390 (*LäSBarT*), 180,012 (*GP2D*), and 543,699 (*8 Sidor*). The distribution of the number of words per superordinate level is presented in Figure 1.

In the first study, we compared the *SUC3* corpus with the *LäSBarT* corpus. To compare

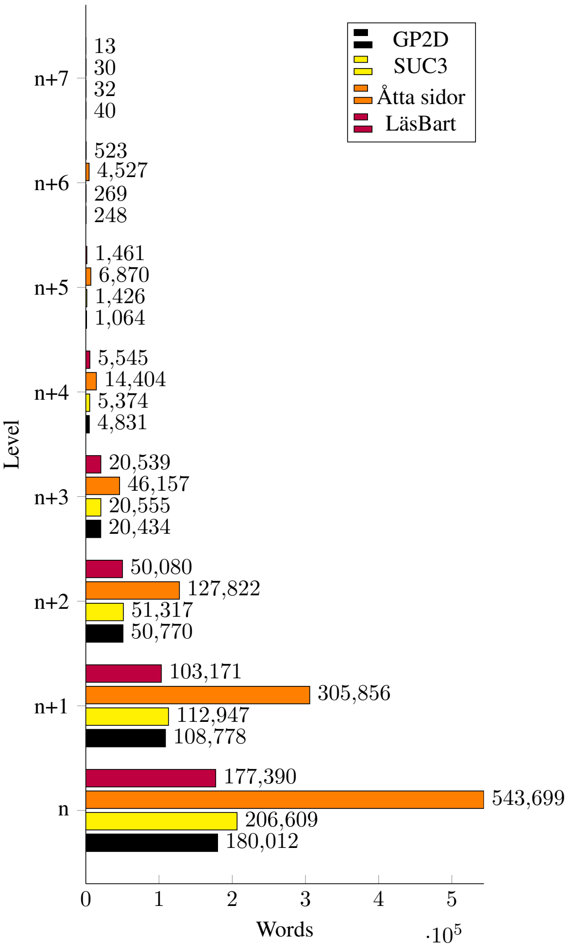

Figure 1: Number of words in the corpora at the various levels. Words at level n are the words in the corpora.

[2] https://spraakbanken.gu.se/verktyg/korp/korpusstatistik

the medians, a Mann-Whitney U test was performed. On average, the words of the *SUC3* corpus had a slightly lower number of steps to the top-level noun $(M = 0.93, Md = 1.0)$ than the words of the *LäSBarT* corpus $(M = 1.02, Md = 1.0)$. This difference was significant $(U = 17489728875.50, n1 = 206,609, n2 = 177,390, p < 0.001, cles = 0.32)$.

In the second study, we compared corpora of the same genre (news texts): *GP2D* and *8 Sidor*. To compare the medians, a Mann-Whitney U test was performed. On average, the words of the *GP2D* corpus had a slightly higher number of steps to the top-level noun $(M = 1.03, Md = 1.0)$ than the words of the *8 Sidor* corpus $(M = 0.93, Md = 1.0)$. This difference was significant $(U = 46166030968.50, n1 = 180,012, n2 = 543,699, p < 0.001, cles = 0.37)$.

The analyses of the relative frequencies of the corpora are presented in Table 1. The words at level n are the words that appear in the corpora[3], and each n+i step refers to the superordinate words. Three of the corpora (*LäSBarT*, *GP2D* and *8 Sidor*) had words represented at the level n+8, but since these words were very few (1, 4 and 1 words respectively), they were excluded from the analysis.

The *SUC3* corpus had the highest relative frequencies at level n+3. The *LäSBarT* corpus had the highest relative frequencies at level n. The *GP2D* corpus had the highest relative frequencies at level n+7. The *8 Sidor* corpus had the highest relative frequencies at level n+3.

All corpora, except for the *LäSBarT* corpus exhibited a tendency of peaking at level n+3 (see Table 1 and Figure 2).

Regarding the news corpora, we can see that the *8 Sidor* corpus has the highest relative frequency at level n, while the highest relative frequency at the standard news corpus *GP2D* is found at level n+4.

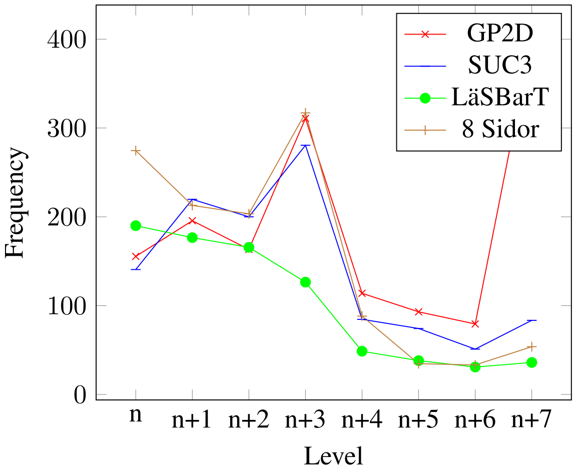

Figure 2: Relative frequencies at each level of the word hierarchy in the corpora.

3.3 Implications for Synonym Replacement Algorithms

From the research on cognitive linguistics referred above, we learnt that basic-level words are more frequently occurring in language, and often monolexemic. Thus, an algorithm shall reward synonym candidates that have high relative frequency and consist of one single lexeme; being monolexemic. To account for the monolexems, information from the frequency corpus about whether or not the word could be interpreted as a compound can be used.

From the corpus analysis, we also found that in the two standard corpora, there seems to be a frequency peak at level n+3. This could be due to the fact that when climbing higher up in the hierarchy of superordinate words, more general words are found, as these words are often more frequently occurring than words with a more specific meaning. When searching for synonyms, we hypothesise that the more general words are not necessarily good synonym candidates. For instance, whereas *horse* can be a good-enough synonym candidate for the word *shetland pony*, the word *animal* might be too general. We conducted experiments with varying levels and chosed to restrict our synonym-seeking algorithm to not go beyond level n+2.

4 Synonym Replacement

Based on the analysis presented in Section 3.3, we developed an algorithm for choosing the best synonym from the extracted nouns and their superordinate words.

[3] We use the notation *level n* to describe the words of the corpora instead of, for example, level 0 words, as we do not know on what level of inclusiveness they actually appear. The words at level n are the words as they appear in the corpora, thus, they could be anywhere on the vertical axis of inclusiveness of the category. The only thing we know is the number of superordinate words, and therefore we chose to use the notation n for the corpus-level and n+i for each superordinate level.

	SUC3	LäSBarT	GP2D	8 Sidor
Level n	140.66	**190.02**	155.44	274.54
Level n+1	219.69	176.59	195.59	212.82
Level n+2	199.97	165.67	163.39	203.38
Level n+3	**280.56**	126.48	310.78	**317.01**
Level n+4	84.60	48.68	113.92	88.22
Level n+5	74.25	38.10	93.04	34.64
Level n+6	51.04	30.88	79.37	33.24
Level n+7	83.47	36.03	**401.41**	53.76

Table 1: Average relative frequencies at each level of the words of the corpora. Highest level frequencies in boldface.

The resulting algorithm is presented in Algorithm 1. It picks, from words at most two levels up in the hierarchy, the most frequent monolexemic word, if such exists, otherwise it picks the most frequent word.

```
Data: candidates: a word chain containing
      the word of the corpus and the
      superordinate words collected from
      Saldo.
Result: best synonym from candidates
candidates.sort(key=frequency);
bestSynonym = candidates[0];
for word in candidates[:3] do
    if word is monolexemic then
        bestSynonym = word;
        break;
    end
end
```

Algorithm 1: The FM algorithm for choosing synonym.

5 Assessment of Synonym Replacement Algorithm

We compared the performance of our combined frequency/monolexemity algorithm (hereafter: *FM*) with two baseline algorithms. The first baseline (*OneLevel*) always chose the word one level higher up in the hierarchy as the best synonym. If there was no superordinate word, the word remained unchanged. The second baseline (*Freq*) always chose the word with the overall highest relative frequency as the best synonym, thus disregarding the monolexemity information.

We ran all algorithms on the nouns extracted from the standard corpora: *SUC3* and *GP2D*.

The results from both corpora regarding number of monolexemic and polylexemic words are presented in Figure 3 and Figure 4 respectively. The relative frequencies after running the algorithms are illustrated in Figure 5.

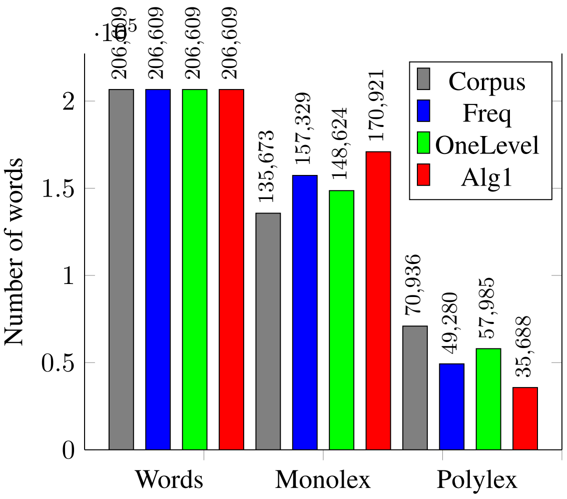

Figure 3: Number of total words, monolexemic words, and polylexemic words in the SUC3 corpus after applying the algorithms. *Corpus* denotes the original values of the specific corpus.

Regarding the *SUC3* corpus, all synonym replacement algorithms increased the number of monolexemic words. The largest increase was observed for the FM algorithm (+35,248), followed by Freq (+21,656), and OneLevel (+12,951). Regarding the relative frequencies, all algorithms increased the average relative frequency of the exchanged words. The largest increase was seen for Freq (+153.68), followed by FM (+120.92), and OneLevel (+34.68).

On the *GP2D* corpus, the number of monolexemic words increased for all algorithms. The largest increase was seen for the FM algorithm (+30,783), followed by the Freq algorithm (+21,091), and OneLevel (+9,482). All synonym

Example word chain	FM	OneLevel	Freq
procent - hundradel - bråkdel - del *percent - centesimal - fraction - part*	procent	hundradel	del
universitet - högskola - skola *university - college - school*	universitet	högskola	universitet
rubel - myntenhet - mynt - pengar *ruble - currency unit - coin - money*	mynt	mynthenhet	mynt

Table 2: Example synonyms chosen by the different algorithms

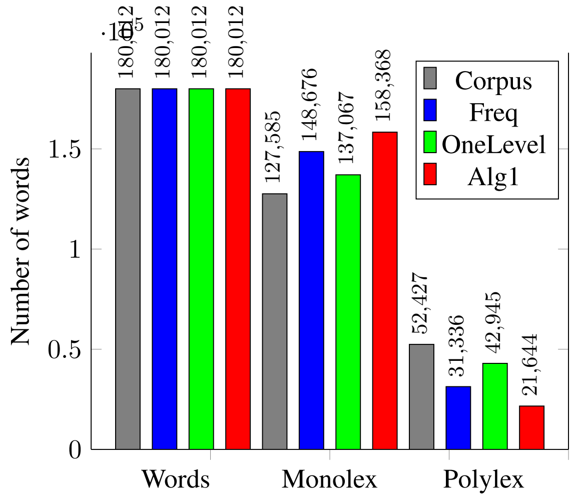

Figure 4: Number of total words, monolexemic words, and polylexemic words in the GP2D corpus after applying the algorithms. *Corpus* denotes the original values of the specific corpus.

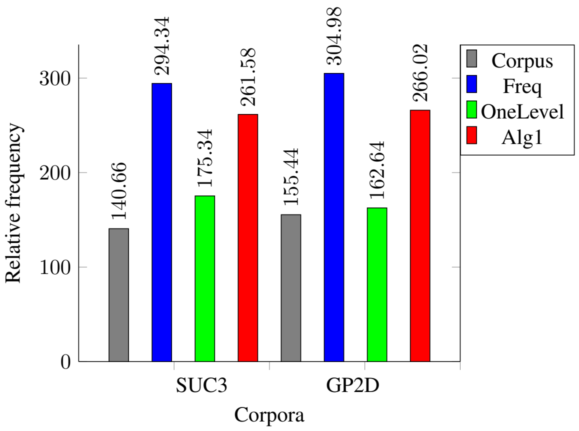

Figure 5: Relative frequencies for each corpus after applying the algorithms. *Corpus* denotes the original values of the specific corpus.

replacement algorithms resulted in a higher average relative frequency, and the largest increase was observed for the Freq algorithm (+149.54), followed by the FM algorithm (+110.58), and OneLevel (+7.2).

Table 2 displays examples of the synonyms chosen by the respective algorithms. As can be seen frequency can sometimes choose a too general word, *del*, whereas OneLevel can pick a too specific word, *myntenhet*.

6 Discussion

The algorithm for finding synonyms proposed in this article is built on theory and corpus studies. This algorithm obviously needs to be evaluated and compared to other methods for extracting synonyms from corpora and lexical resources. It would be valuable to compare the algorithm with synonyms from, for example, the SynLex lexicon, and to evaluate whether the exchanged synonyms are simpler, when consulting lexicons of base vocabularies, as well as humans. It can also be enhanced with techniques to utilise semantic and synonym similarity (Kann and Rosell, 2005).

The corpus analyses were not conclusive, and, although further analyses will probably not present results that argues against the proposed algorithm, further investigations may be important for the study of language use and we therefore present a more detailed discussion on the corpus study.

We hypothesised that simple texts would exhibit a tendency towards the use of more basic-level words, when compared with texts written in standard Swedish. However, there was no clear support for this hypothesis. In the statistical analysis, we compared very large samples, and the presence of statistical significance is not surprising. When comparing the means and medians of the datasets, it is clear that the differences are small and the results should be interpreted with caution.

The results of the first study revealed that the *SUC3* corpus had a significantly lower average number of steps to the top-level noun, than the *LäSBarT* corpus. Since our hypothesis was that the texts of the corpus of simple text would have a lower average number of steps to the top-level noun, these results showed a difference in the opposite direction.

The second study was normalised for genre, in the sense that the compared corpora contained texts of the same genre. The simple news corpus *8 Sidor* had a significantly lower number of steps to the top-level noun than the standard news corpus *GP2D*. This tendency is further supported by the results of the relative frequency analysis, where we clearly see that the *8 Sidor* corpus has relatively high average relative frequency at the base level (level n), although exhibiting the highest frequencies at level n+3, whereas the *GP2D* corpus generally had lower average frequencies at level n and the highest frequencies at level n+7.

Regarding the analyses of the relative frequencies, we would expect the standard corpora to have lower relative frequencies at the base level (level n) than the corpora of simple text. This difference can be observed in the *LäSBarT* corpus, which had the highest relative frequency scores at level n, but is less prominent in the *8 Sidor* corpus. However, even if the *8 Sidor* corpus exhibits the highest relative frequencies at level n+3, it is noteworthy that the frequencies are relatively high even at the lower levels. The level n score is the second highest frequency score for this corpus, and much higher when compared to the level n score of the standard corpus of the same genre, *GP2D*.

The *GP2D* corpus had the highest average frequency at level n+7, indicating that the words used in this corpus are more specific than in the other corpora. However, it should be noted that this high relative frequency score is based on a relatively low number of words (40), and that this corpus also exhibit the frequency peak at level n+3.

For *SUC3* and *8 Sidor*, the most frequent words are found at level n+3. This would mean that the more basic-level nouns could be found if we choose the superordinate words three levels above the original word. However, it could also indicate that the words at this level are higher up at Rosch's vertical axis, thus being more inclusive than the basic-level words, and therefore more frequent (compare: *shetland pony*, *horse*, *animal*).

When designing this study, we made a number of assumptions that can be discussed, such as the assumption of the nature of texts in simple Swedish versus texts in standard Swedish. We made the assumption, according to Rosch's claims of basic-level terms, that the proportion of such constructions would be higher in the simple corpora. This assumption should be tested, for example by counting the relative frequencies of some base vocabulary list words (Heimann Mühlenbock and Johansson Kokkinakis, 2012) in both corpora.

The usage-based thesis of cognitive linguistics implies that we gain knowledge about the linguistic system by studying authentic language in use. To this background, it seems reasonable that a corpus study would be suitable for studying linguistic phenomena. However, there are some drawbacks of using such methods. One of the problems is that we worked with four very different corpora. Can we really say that a corpus reflects authentic and direct language use? For example, one commonly mentioned measure in this context is frequency. A frequency measure can provide information on how commonly used certain linguistic constructions are. However, what we see clearly in this study is that if we compare corpora of different characteristics, the frequency measures will differ between corpora depending on text type. A corpus of medical texts will have frequent constructions that do not even exist in a corpus of children's literature. The same issue will probably be manifested if we compare texts of different linguistic activities, such as spoken language with written language. This means that the insights that we can draw of the cognitive processes underlying the studied linguistic phenomenon will be very specific to the kind of corpus that we study. To compare corpora, we must make sure that the corpora are comparable, and consider the factor of language use reflected in the texts of the corpora when generalising our findings to a larger context.

7 Conclusion

The aim of this paper was to develop an algorithm for synonym replacement based on theories of basic-level nouns. We also presented results from a study exploring whether corpora of simple texts contain a higher proportion of basic-level nouns than corpora of standard Swedish, and to see how this could be used in the context of lexical text simplification.

We observed that the corpus of simple news text did indeed include more basic-level nouns than the corpus of standard news. This in turn shows that lexical simplification, through the use of base-level nouns, may benefit from traversing a word hierarchy upwards. This could serve as a complement to the often-used replacement methods that rely on word length and word frequency measures.

We presented techniques for finding the best synonym candidate in a given word hierarchy, based on information about relative frequencies and monolexemity. We saw that all synonym replacement techniques, including the baseline methods, increased the number of monolexemic words and relative frequencies. The FM algorithm aimed to reward high relative frequencies and monolexemity, while not climbing the word hierarchy too high, and seems to perform well with respect to these criteria. Future work includes further evaluation of this algorithm, and comparison to other synonym replacement strategies.

References

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, Department of Swedish, University of Gothenburg, Gothenburg, Sweden.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248. ACM.

Gianni Barlacchi and Sara Tonelli. 2013. Ernesta: A sentence simplification tool for children's stories in italian. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 476–487.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501.

Lars Borin, Marcus Forsberg, and Lennart Lönngren. 2008. SALDO 1.0 (Svenskt associationslexikon version 2). *Språkbanken, Göteborgs universitet.*

Lars Borin and Markus Forsberg. 2014. Swesaurus; or, The Frankenstein approach to Wordnet construction. In *Proceedings of the Seventh Global Wordnet Conference*, pages 215–223.

Arnaldo Candido Jr, Erick Maziero, Caroline Gasperin, Thiago AS Pardo, Lucia Specia, and Sandra M Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42. Association for Computational Linguistics.

Yvonne Canning and John Tait. 1999. Syntactic simplification of newspaper text for aphasic readers. In *ACM SIGIR'99 Workshop on Customised Information Delivery.*

John Carroll, 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*, volume 1, pages 7–10. Citeseer.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and Methods for Text Simplification. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96).*

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

Walter Daelemans, Anja Höthker, and Erik F Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC).*

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Siobhan Devlin and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 225–226.

Biljana Drndarević and Horacio Saggion. 2012. Towards automatic lexical simplification in Spanish: an empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 8–16.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm Umeå Corpus version 2.0.

Vyvyan Evans. 2019. *Cognitive Linguistics(2nd edition).* Edinburgh: Edinburgh University Press.

Katarina Heimann Mühlenbock and Sofie Johansson Kokkinakis. 2012. SweVoc - a Swedish vocabulary resource for CALL. In *Proceedings of the*

SLTC 2012 workshop on NLP for CALL, pages 28–34, Lund. Linköping University Electronic Press.

Firas Hmida, Mokhtar B. Billami, Thomas François, and Núria Gala. 2018. Assisted lexical simplification for French native children with reading difficulties. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 21–28, Tilburg, the Netherlands. Association for Computational Linguistics.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16. Association for Computational Linguistics.

Viggo Kann and Magnus Rosell. 2005. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference*, pages 105–110, Stockholm.

MTM. 2020. Att skriva lättläst. `https://www.mtm.se/var-verksamhet/lattlast/att-skriva-lattlast/`. Accessed: 2020-10-05.

Katarina Mühlenbock. 2008. Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. In *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8 of *Acta Universitatis Upsaliensis*, pages 327–329, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.d. thesis, University of Sheffield, Sheffield, UK.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-braem. 1976. Basic objects in natural categories. *Cognitive Psychology*.