

Auxiliary Techniques to Help Readers Understand Texts

Daniel Holmer, Arne Jönsson

Department of Computer and Information Science
Linköping University, Sweden
{daniel.holmer, arne.jonsson}@liu.se

Abstract

We explore three auxiliary techniques for automatic text adaptation (ATA)—epithets for nouns, explanations for keywords, and syllabification—to aid reading for individuals with reading difficulties, specifically targeting students with dyslexia and intellectual disabilities (ID). The methods are assessed as part of a broader investigation into ATA, with the goal of making texts more accessible to readers with varying needs. For an initial evaluation of the three techniques, we conduct a study with readers with average reading skills. The evaluation indicates that while all three techniques perform well in terms of accuracy, their usefulness varies. Epithets were found to be less helpful, possibly because they can add unnecessary information for readers with average reading skills, though they might benefit individuals with ID. Keyword explanations were generally helpful and correct, with the caveat that inaccuracies can occur, particularly with rare or domain-specific terms. Syllabification's effectiveness was found to depend on the word being processed. We conclude that while these techniques show promise, their utility may vary among different reader groups, and further research is needed to assess their impact on readers with specific reading difficulties.

Keywords: Text adaptation, Epithet, Word explanation, Syllabification

1. Introduction

Text adaptation normally includes lexical simplification, syntactic simplification and various forms of text summarisation. Other important techniques to make texts easier to read include font type and size, line width and line spacing. But there are other means to make texts easier to read that are in between the former language technology techniques that rewrite a text and the latter more surface oriented, requiring no linguistic processing. In this paper we present three such techniques: the use of epithets to help understand certain nouns, explaining central keywords in a text instead of simplifying them, and splitting words into syllables.

The techniques are part of a larger investigation on automatic text adaptation (ATA) techniques for people with reading problems. The larger project aims to adapt texts to suit the needs of students with dyslexia and to students with intellectual disabilities (ID). Both reader groups have been previously acknowledged within the ATA field, but the reading impairments have partly different origins, and the individuals within each group experience varying difficulties.

For instance, dyslexia is associated with issues of establishing the grapheme-phoneme correspondences needed for decoding (Vellutino et al., 2004), which causes specific problems with word decoding. As decoding is cognitively demanding, the performance of simultaneous mental processes (Høien and Lundberg, 2013) is affected. Examples of reading challenges associated with dyslexia are long words and low-

frequency words (Rello et al., 2013b), homophones, words that are orthographically similar, new words, and non-words (Rello et al., 2013a). Individuals with dyslexia also have limited ability to make various types of inferences (Simmons and Singleton, 2000) as most of the cognitive resources are used for decoding which will affect reading comprehension negatively.

The reading difficulties of individuals with ID are reflected in the decoding and reading comprehension skills (Nilsson et al., 2021a,b). There are limitations in several cognitive abilities, such as working memory and executive functions (Danielsen et al., 2012), and impaired language skills are related to vocabulary and grammatical understanding (Nilsson et al., 2021a). Therefore reading comprehension is often severely compromised and more than half of the population in the age group 11 read words letter by letter, which is characteristic of the early phases of learning to read in a typical population (Lemons et al., 2013).

We can, thus, conclude that reading difficulties vary and that the usefulness of different techniques may vary, and that there can be incentive to try a variety of ATA techniques.

2. Techniques

In the main study we will adapt texts using automatic text summarisation and text simplification. In this paper we focus on three techniques that may have a potential to further facilitate reading for our target groups, as well as other groups with

reading problems, namely epithets, key word explanations and syllabification. We select these three techniques based on the guidelines developed by the Swedish Agency for Accessible Media (MTM, 2021) and various studies, c.f. (Kearns and Whaley, 2019). These guidelines suggest different ways to write texts in an accessible way, for instance regarding linguistic constructions and to select simple and short words. However, in some instances there are no suitable simple and short synonyms to a word. Difficult words should therefore be given an explanation (MTM, 2021). Two of our techniques aim to provide this in different ways; the first by providing short, descriptive epithets to give more context to a word, and the second by providing a clarifying explanation to certain keywords.

2.1. Data

In all studies we will use three texts on minority languages in Sweden, one on Yiddish, one on Finnish and one on Swedish Sign Language. The texts are a part of education material provided by The Institute for Language and Folklore (Isof)¹, where texts on all of Sweden's six minority languages are covered. In our selection of the three texts, we ran an analysis of six different text complexity metrics on all texts, and selected the three texts that had the most similar complexity according to the metrics. The metrics were selected to provide information about different aspects of the texts. we use LIX (Björnsson, 1968) as a surface metric regarding sentence and word length, OVIX (Hultman and Westman, 1977) for idea density, three syntactic metrics (AVG_DEP_DISTANCE_DEPENDENT, AVG_SENTENCE_DEPTH, and NOMINAL_RATIO) (Falkenjack, 2018), and the cohesion metric ADJACENT ANAPHORS².

2.2. Epithets

Epithets are descriptive terms accompanying the name of a person, place, or thing. For epithets we use a pipeline of two BERT-models. The first is fine-tuned for named entity recognition (NER)³. The model is trained to identify different types of entities, for instance persons, locations and organisations. For each such identified entity, we add a [MASK] token in the position before the entity. We

¹<https://www.isof.se/nationella-minoritetsprak/loromedel/loromedel-fran-isof>

²Index 38 from the Coh-Metrix documentation found at http://cohmetrix.memphis.edu/cohmetrixhome/documentation_indices.html

³<https://huggingface.co/KBLab/bert-base-swedish-lowermix-reallysimple-ner>

then feed the whole sentence to a second BERT-model (Malmsten et al., 2020)⁴, which is tasked to predict the [MASK]-token. In essence, this mimics the MLM pre-training step described in (Devlin et al., 2019). We add an additional post processing step that cross references the predicted epithet token to a list of manually curated epithets, to make sure that added tokens are a theoretically valid epithet. A typical epithet to, for instance, the word Sweden is "the country" producing "the country Sweden".

2.3. Keywords

To extract keywords we use a system based on YAKE! (Campos et al., 2020), a custom n-gram extractor, and KeyBERT (Grootendorst, 2020).

We use YAKE! and the n-gram extractor to find possible keyword candidates. These candidates are then fed to KeyBERT, which ranks the most relevant keywords from the candidate list. KeyBERT follows an approach where it uses embeddings from a BERT-model in two steps. First, it works on the word level, where an embedding for each candidate keyword are created. Second, it creates embeddings on the document level. To select the most important keywords, the cosine-similarities between all the candidate and document embeddings are calculated, and the keywords with the highest similarity to the documents are considered to be the most relevant.

While it is possible to let KeyBERT treat the entire text as keyword candidates, we opted for the pre-processing approach where YAKE! and the n-gram extractor provide a limited selection of candidates. The reason for this is two-fold; we want to have greater control over what words are possible for selection (we select proper nouns, adjectives, and nouns as valid candidates for the n-gram extractor), and due to limited hardware we want to avoid creating BERT-embeddings for every individual word in the whole text. To further alleviate the computational need of KeyBERT, we use a distilled version of the Swedish SBERT model (Rekathati, 2021) from KBLab⁵.

The identified keywords are then given an explanation by prompting the LLM GPT4-TURBO-PREVIEW⁶ from OpenAI. We use a zero-shot prompt where the model is instructed to explain the given word in a simple way, and avoid using difficult words:

⁴<https://huggingface.co/KBLab/bert-base-swedish-cased>

⁵<https://huggingface.co/KBLab/sentence-bert-swedish-cased>

⁶All further references to GPT-4 are the state of GPT4-TURBO-PREVIEW as accessed from OpenAI:s API in February 2024.

Provide an explanation in no more than one sentence for this word: **{word}**. The explanation should be easy to understand and not contain long or difficult words. Use words that are easy to understand.⁷

where **{word}** is the given keyword to be explained.

We keep the hyperparameters at their default values in the CHAT-COMPLETIONS interface from OpenAI.

A typical example is *Teckenspråk – Teckenspråk är ett språk där man använder händerna, ansiktsuttryck och kroppsrörelser för att kommunicera istället för att prata med rösten.* (eng. Sign Language – Sign language is a language where hands, facial expressions, and body movements are used to communicate instead of speaking with the voice.)

2.4. Syllabification

The syllabification technique used in this research is based on morphological rather than phonetic principles. For syllabification the texts are first converted to plain text using docx2txt⁸.

We then use the compound analysis of the Sparv pipeline⁹ where tokens and their POS tags are looked up in the SALDO lexicon (Borin et al., 2013) and enriched with compound information.

The compound analysis includes identifying candidate words according to criteria such as having a prefix in the SALDO lexicon, being compound, having a suffix with certain properties such as being noun, verb or adjective, etc. The candidates are then ranked based on criteria such as number of compounds and a statistical model¹⁰ (Borin et al., 2016).

The three texts contain a total of 1788 words of which 254 are syllabified. Of those 254 words there are 200 unique instances.

The results from the syllabification often affects compound words, such as, *väst•finska* (eng. west•Finnish), but there are also instances of syllabification's inside words, like *ar•bete* (eng. work).

⁷The full Swedish prompt is constructed as: "Ge en förklaring på max en mening till detta ord: **{word}**. Förklaringen ska vara lätt att förstå och inte innehålla långa eller svåra ord. Använd ord som är lätta att förstå."

⁸<https://pypi.org/project/docx2txt/>

⁹<https://spraakbanken.gu.se/sparv/#/user-manual/available-analyses?id=compound-analysis-with-saldo>

¹⁰Details in Swedish on <https://spraakbanken.gu.se/faq/hur-fungerar-sparvs-sammansattningsanalys>

3. Evaluation

The three techniques have been assessed by readers with average reading skills, 10 students and teachers at Linköping University, in a survey comprising 10 random instances of each technique. The instances are taken from the three texts that are to be used in the main study, see Section 2. The survey uses a five grade Likert scale with the following items:

Helt enig (eng. Strongly agree), grade 5

Delvis enig (eng. Somewhat agree), grade 4

Neutral, grade 3,

Delvis oenig (eng. Somewhat disagree), grade 2

Helt oenig (eng. Strongly disagree), grade 1

For the ten epithets and the ten keyword explanations two questions were asked:

- *epitetet|ordförklaringen är korrekt* (eng. the epithet|word explanation is correct)
- *epitetet|ordförklaringen underlättar förståelsen* (eng. the epithet|word explanation facilitates understanding)

For the syllabifications the first question was instead formulated as if the syllabification is good (Sw. *bra*), as there is not always an obvious correct syllabification and the second was formulated as *Orduppdelningen underlättar läsningen* (eng. The syllabification facilitates reading) as syllabification is more an aid for reading.

Table 1 shows the results from the survey. We present both results interpreting the Likert scale as an interval scale, mean and standard deviation, as well as an ordinal scale, median. As can be seen from Table 1 all techniques perform well, median 5.0. Looking at the mean we see that there are some deviations and when we further study the various items we can identify some interesting patterns.

Looking at the epithets we see that some of them are considered less correct and helpful. For instance, words with the epithet "state" (e.g. the state of Sweden) as opposed to the epithet "country" (e.g. the country Sweden) are considered less correct and also considered less helpful. See Figure 1.

Overall epithets facilitate understanding the least, median only 3.0.

Word explanations are also correct and much more useful. Interestingly, one word explanation is wrong, the explanation about the Swedish so called *Judereglementet* (eng. 'The Jewish Regulations') explains rules of the sport Judo. This is also

	Median	Mean	Standard deviation
The epithet is correct	5.0	4.15	1.445
The epithet facilitates understanding	3.0	2.99	1.507
The word explanation is correct	5.0	4.31	1.309
The word explanation facilitates understanding	5.0	4.25	1.268
The syllabification is good	5.0	4.27	1.318
The syllabification facilitates reading	4.0	3.67	1.537

Table 1: Results from the evaluation

2. landet Sverige (0 poäng)

[Mer information](#)



10. staten Sverige (0 poäng)

[Mer information](#)

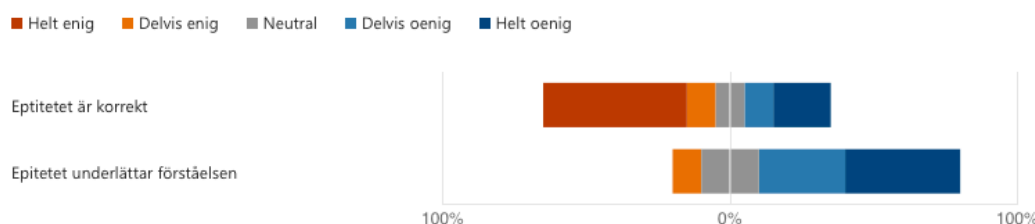


Figure 1: Two extracts from the analysis of epithets. They only differ in epithet, "the country Sweden", top, and "the state Sweden", bottom. The first item translates to "the epithet is correct" and the second "the epithet facilitates understanding". The various grades are translated in the bullet list above.

observed by more or less all participants and considered both not correct and not to facilitate understanding. If we remove this item we get the results shown in Table 2 where the mean and standard deviation clearly shows that word explanations are both correct and facilitates understanding. Even for readers with average reading skills.

Syllabifications, finally, are also considered correct but does not facilitates reading as much. Here we see differences between the two types of syllabifications that the technique provides, one that more or less divides Swedish compounds into their parts, such as *tecken•språket* (eng. the sign language) and the other that divides words into syllables, such as *as•kan* (eng. the ashes). The latter is regarded less correct and much less useful. This also seems to depend on the length of the word, short words such as *askan* are considered both

not correct and not to facilitate reading whereas slightly longer words, such as *ar•bete* (eng. work) are considered less incorrect and not as bad when it comes to facilitate reading. Figure 2 clearly illustrates this, where readers with average reading skills rate compounds high, bottom, but not short words.

4. Discussion

All techniques perform well and we can conclude that it is possible to accurately add epithets to nouns, explain keywords, and perform syllabification on words. Their estimated usefulness, for readers with average reading skills, varies, however. Estimated usefulness does not necessarily mean that texts with these features would have

	Median	Mean	Standard deviation
The word explanation is correct	5.0	4.6	0.903
The word explanation facilitates understanding	5.0	4.5	0.946

Table 2: Revised word explanations results

22. arbete (0 poäng)

[Mer information](#)

■ Helt enig ■ Delvis enig ■ Neutral ■ Delvis oenig ■ Helt oenig



23. tecken•språket (0 poäng)

[Mer information](#)

■ Helt enig ■ Delvis enig ■ Neutral ■ Delvis oenig ■ Helt oenig

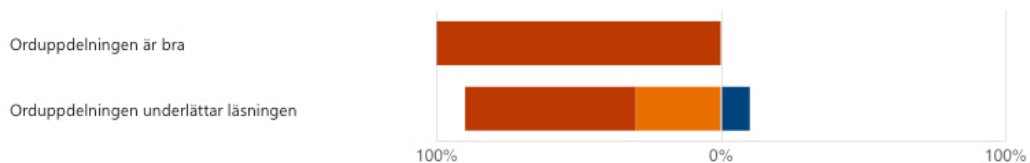


Figure 2: Two extracts from the analysis of syllabifications, *arbete*, (eng. work) top and *tecken•språket* (eng. the sign language), bottom. The first item translates to "the syllabification is good" and the second to "The syllabification facilitates reading". The various grades are translated in the bullet list above.

been helpful for the readers in the study. Rather, they can see potential gains from using these features.

Epithets clearly do not always facilitate understanding even if they are correct. This may not be surprising for readers with average reading skills, where the epithet can be seen as adding a word that is not necessary to understand the word. We believe that the same is true for people with dyslexia, but for people with intellectual disabilities we believe that it may be useful.

Recently, different types of LLMs have shown great results on many NLP-tasks, and in particular generation tasks. In the case of word explanations, it is clear that an LLM in the like of GPT-4 can provide more helpful explanations than previous techniques. However, in the task of generating an epithet before identified keywords, the advantages of such LLMs are not as obvious. In our experiments, we also generated epithets using LLMs by prompting GPT-4 to adhere to the experimental settings of our BERT-based system.

The gains of the more environmentally expensive GPT-4 model are slight. The main difference can be seen in the cases like described in 1, where our BERT-based system sometimes would put the epithet "state" instead of "country" in conjunction to countries. When prompting GPT-4, it consistently delivers "country". Otherwise, the results are shown to be nearly identical. It is however possible that the advantages of an LLM-based solution would be more evident where even more complex words would have to be associated with an epithet, or in a setting where for example phrases of epithets would be allowed (i.e. "the Nordic country Sweden", or "the American city New York"). This is something we will investigate in further studies.

Word explanations are considered more helpful. When they are correct they also facilitate understanding. However, there is a risk that they are wrong. As expected, large language models (in this case GPT-4), struggle to provide a feasible explanation for rare and highly domain specific words. For instance, this is demonstrated in the earlier

example, where the term *Judereglementet* (eng. 'The Jewish Regulations') resulted in an explanation of the rules of the sport Judo. It is reasonable to believe that the term was not all that common in the model's training data, and therefore these kinds of hallucinations might appear.

Syllabification is the technique that most depends on which word that is being processed. Many of the words were considered less useful, and sometimes not even considered correct by readers with average reading skills. We believe, however, that syllabification of these words may well help people with dyslexia, not only compounds but also words that are not compounds. Unless they are very short. In our further studies we will not syllabify words shorter than 6 characters, c.f. (Björnsson, 1968). For our three texts, see Section 2 we are then left with 215 words to be syllabified.

5. Summary

In this paper, we have presented results from an investigation of three techniques that could be used in conjunction with text summarization and text simplification to facilitate reading for different target groups. The three techniques are; the addition of epithets, explanations of keywords, and splitting words into syllables. The techniques were evaluated in a survey taken by readers with average reading skills. The evaluation is small, and the results are merely indicative, but they show that all three techniques are mostly performing well on their respective tasks. The usefulness for readers with average reading skills are, however, not always apparent. For epithets it very much depends on the epithet, some epithets are considered helpful, but many epithets do not facilitate understanding. Word explanations on the other hand are consistently believed to be helpful, while the syllabification is also highly dependant on what word that is syllabified. In further studies we will investigate how these techniques are perceived by readers with different reading difficulties.

6. Bibliographical References

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm, Sweden.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology*

Conference, Umeå University, 17-18 November, 2016.

Lars Borin, Markus Forsberg, and Lennart Lönngrén. 2013. SALDO: a touch of yin to WordNet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Henrik Danielsson, Lucy Henry, David Messer, and Jerker Rönnerberg. 2012. Strengths and weaknesses in executive functioning in children with intellectual disability. *Research in developmental disabilities*, 33(2):600–607.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Johan Falkenjack. 2018. *Towards a model of general text complexity for Swedish*. Licentiate thesis, Linköping University Electronic Press.

Maarten Grootendorst. 2020. [Keybert: Minimal keyword extraction with bert](#).

Torleiv Høien and Ingvar Lundberg. 2013. *Dyslexi - från teori till praktik, 2.*, [uppda edition. Natur & Kultur, Stockholm.

Tor G. Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. LiberLäromedel, Lund.

Devin M. Kearns and Victoria M. Whaley. 2019. Helping students with dyslexia read long words: Using syllables and morphemes. *Teaching Exceptional Children*, 51(3).

Christopher J. Lemons, Naomi Zigmund, Amanda M. Kloof, David R. Hill, Alicia A. Mrachko, Matthew F. Pattera, Thomas J. Bosta, and Shawn M. Davis. 2013. [Performance of students with significant cognitive disabilities on early-grade curriculum-based measures of word and passage reading fluency](#). *Exceptional Children*, 79(4):408–426.

Martin Malmsten, Love Börjeson, and Chris Hafenden. 2020. [Playing with words at the national library of Sweden – making a Swedish BERT](#). *arXiv preprint arXiv:2007.01658*.

MTM. 2021. Att skriva lättläst. <https://www.mtm.se/var-verksamhet/lattlast/att-skriva-lattlast/>. Accessed: 2021-10-05.

Karin Nilsson, Henrik Danielsson, Åsa Elwér, David Messer, Lucy Henry, and Stefan Samuelsson. 2021a. [Decoding Abilities in Adolescents with Intellectual Disabilities: The Contribution of Cognition, Language, and Home Literacy](#). *Journal of Cognition*, 4(1):1–16.

Karin Nilsson, Henrik Danielsson, Åsa Elwér, David Messer, Lucy Henry, and Stefan Samuelsson. 2021b. [Investigating reading comprehension in adolescents with intellectual disabilities: Evaluating the simple view of reading](#). *Journal of Cognition*, 4(1):1–16.

Faton Rekathati. 2021. The KBLab blog: Introducing a Swedish sentence transformer. <https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/>.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer.

Fiona Simmons and Chris Singleton. 2000. The Reading Comprehension Abilities of Dyslexic Students in Higher Education. *Dyslexia*, 6:178–192.

Frank R Vellutino, Jack M Fletcher, Margaret J Snowling, and Donna M Scanlon. 2004. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40.