1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 6–13
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Is it simpler? An Evaluation of an Aligned Corpus of Standard-Simple Sentences

Evelina Rennes
Department of Computer and Information Science, Linköping University
RISE, Research Institutes of Sweden
Linköping
evelina.rennes@liu.se

Abstract

Parallel monolingual resources are imperative for data-driven sentence simplification research. We present the work of aligning, at the sentence level, a corpus of all Swedish public authorities and municipalities web texts in standard and simple Swedish. We compare the performance of three alignment algorithms used for similar work in English (Average Alignment, Maximum Alignment, and Hungarian Alignment), and the best-performing algorithm is used to create a resource of 15,433 unique sentence pairs. We evaluate the resulting corpus using a set of features that has proven to predict text complexity of Swedish texts. The results show that the sentences of the simple sub-corpus are indeed less complex than the sentences of the standard part of the corpus, according to many of the text complexity measures.

Keywords: parallel corpus, monolingual alignment, automatic text simplification, text complexity

1. Introduction

Automatic Text Simplification (ATS) denotes the process of transforming a text, semantically, syntactically or lexically, in order to make it easier while preserving meaning and grammaticality. The simplification of text can have different purposes. Historically, it has been used as a preprocessing step to facilitate other natural language processing tasks, such as machine translation and text summarisation. The intuition was that a simpler syntactic structure of input texts would lead to less ambiguity, which would improve text processing performance.

Another purpose of ATS is to make texts available to a broader audience, for example by adapting texts for people with different kinds of reading difficulties (Saggion, 2017). Examples of target groups that have been accounted for within the field are people with dyslexia, people with aphasia, children, the deaf and hearing-impaired, second language learners, and the elderly.

Data-driven techniques have gained ground the last years within the field of natural language processing, and the simplification field is no exception. Recent approaches regard simplification as a task analogous to (monolingual) machine translation (Specia, 2010; Coster and Kauchak, 2011b; Coster and Kauchak, 2011a; Wubben et al., 2012; Xu et al., 2016; Nisioi et al., 2017; Zhang and Lapata, 2017; Zhang et al., 2017).

One well-recognised issue with data-driven techniques is that these techniques typically demand large-scale high-quality data resources, which can be problematic for less-resourced languages. A widely used resource in previous automatic text simplification research is Wikipedia and Simple English Wikipedia (Zhu et al., 2010; Coster and Kauchak, 2011b; Hwang et al., 2015; Kajiwara and Komachi, 2016), but its quality as a resource has been questioned (Xu et al., 2015). The collaborative and uncontrolled nature of Wikipedia makes it somewhat unreliable as a resource, and the authors pointed out that simple articles generally are not rewritten versions of the standard articles, which can be problematic when attempting to perform sentence alignment.

Another commonly used resource is the Newsela corpus[1]. Newsela contains 1,130 original news articles in English, manually simplified to 3–4 complexity levels by professional writers. The readability levels correspond to education grade levels, thus targeting children of different reading levels. Although there are many advantages of Newsela, such as the high quality of the texts, there is one disadvantage: researchers are not allowed to publicly release model output based on this corpus, which in turn hinders model comparison. The Newsela corpus has been used in some studies for text simplification (Zhang and Lapata, 2017; Alva-Manchego et al., 2017; Scarton et al., 2018).

The need for more and better resources for sentence simplification was highlighted by Alva-Manchego et al. (2020), and proposed as one of the key topics that should be addressed by the field.

In Sweden, most websites of public authorities and municipalities have versions adapted to people in need of simple text. These texts are often based on guidelines learned from the professional experience of expert writers and editors. The Swedish Agency for Accessible Media (MTM) describes some of these guidelines[2]:

- The text should be adapted to the type of reader that will read the text
- The text should have a common thread and capture the interest of the reader immediately
- The context should be clear, and the text should not demand any extensive prerequisites

[1] https://newsela.com/data

[2] https://www.mtm.se/produkter-och-tjanster/lattlast/om-latta-texter/

- The text should contain everyday words and the text rows should be short
- If a picture is presented next to a text, it should interplay with the text
- The language and presentation should be adapted to the specific demands and purposes of the specific type of media

These properties are, for obvious reasons, difficult to model in a concrete and unambiguous way to be fed into a system that automatically simplifies text.
Professionally written texts comprise, however, concrete examples of sentences that adhere to these guidelines. They can therefore be used for learning how experts write simple text. This motivated us to collect a corpus of web texts from Swedish public authorities and municipalities (Rennes and Jönsson, 2016).
The collected corpus contained a total of 1,629 pages in simple Swedish, and 136,501 pages in standard Swedish, with a total of 29.6 million tokens.
The corpus was aligned using three different alignment algorithms, broadly following Kajiwara and Komachi (2016). The alignment algorithms, originally proposed by Song and Roth (2015); Average Alignment (AA), Maximum Alignment (MA), and Hungarian Alignment (HA), align sentence pairs by calculating and combining the similarities of word embeddings to create a sentence similarity score.
The AA algorithm bases the sentence similarity on the average of the pairwise word similarities of all words of a pair of sentences. The MA algorithm considers the word pairs that maximise the word similarity of all words of a pair of sentences, and the sentence similarity score is given by the sum of the word similarity scores. The HA algorithm determines the sentence similarity by calculating the lowest cost (in our case, the highest cosine value) for every possible word pair, and the resulting sum is normalised by the length of the shortest sentence in the sentence pair.
Thus, for all algorithms, we could alter the *word similarity threshold* (the threshold of when a word pair is regarded similar enough) and *sentence similarity threshold* (the threshold of when a sentence pair is similar enough and should be aligned).
A few modifications of the Kajiwara and Komachi (2016) implementation were made. The language was changed to Swedish, and unknown words, so called Out-of-Vocabulary (OOV) words, were treated differently. Since Kajiwara and Komachi (2016) used word embeddings trained on a large-scale corpus, they disregarded the OOV words when calculating the sentence similarity scores. However, since we used a much smaller set of Swedish word embeddings, Swectors (Fallgren et al., 2016), ignoring OOV words was not a viable approach. Instead, we used Mimick (Pinter et al., 2017) to train a recurrent neural network at the character level, in order to predict OOV word vectors based on a word's spelling. Mimick works by generating approximated word embeddings for OOV words. The intuition behind this approach is that word embeddings that are generated based on the spelling of a word provide a better vector estimation than other common methods (such as creating a randomised word embedding) since they capture features related to the shape of a word.
In this article, we present detailed results on the nature of the different algorithms using a combination of evaluations. In Section 2.1., we investigate at what sentence similarity threshold humans perceive the aligned sentence pairs as semantically similar. In Section 2.2., we aim to find the algorithm and the best combination of parameters to maximise alignment performance. In Section 2.3., we investigate whether the sentences in the aligned sentence pairs differ in complexity. In Section 3., results and methodological considerations are discussed, and the conclusions are presented in Section 4..
The main contribution of this work is the provision and evaluation of a new text simplification corpus for Swedish.

2. Evaluations

A total of three evaluations were performed. The first two evaluations aimed to tune the values of the word and sentence similarity thresholds to maximise the performance of the algorithms. An aligned corpus was then created of sentence pairs using the best-performing threshold values.
The third evaluation aimed to investigate whether the aligned corpus consisted of sentence pairs that differed in complexity, i.e. if we really had a corpus of standard and simple Swedish. Since the sentences are extracted from corpora consisting of standard and simple documents, it is intuitive that the extracted sentences are good representatives of standard and simple text segments. However, given the way the corpus was created, we cannot know that the sentence pairs are true alignments, that is, that the simple sentence is a *simplified version of the standard sentence*. The third evaluation aims to investigate whether the sentences of the different parts of the corpus in fact differs in complexity.

2.1. Evaluation I: Human Evaluation

The quality of the sentence pairs generated by the alignment algorithms was evaluated in a human evaluation conducted through a web survey. The word threshold value was set to 0.49 following Kajiwara and Komachi (2016). The intuition behind this evaluation was to see at what sentence threshold humans perceive the aligned sentences as semantically similar.

2.1.1. Procedure

From the three corpora generated by the different algorithms, we randomly picked three sentence pairs per similarity interval (0.51–0.60, 0.61–0.70, 0.71–0.80, 0.81–0.90, 0.91–1.0). The number of sentence pairs aligned by the AA algorithm were, however, very few ($<$10). AA was therefore excluded from this evaluation. For MA and HA a total of 30 sentence pairs were extracted.
All extracted pairs from HA and MA were then included in a web survey, and participants were asked to grade the sentence pairs on a four-graded scale regarding similarity. The grading was based on categories previously used to create a manually annotated data set (Hwang et al., 2015). For this evaluation, the categories were translated into Swedish and slightly reformulated to suit non-experts. The reformulated categories were:

1. **Meningarna handlar om helt olika saker**
   The sentences treat completely different things

2. **Meningarna handlar om olika saker men delar en kortare fras**
   The sentences treat different things, but share a shorter phrase

3. **En menings innehåll täcks helt av den andra meningen, men innehåller även ytterligare information**
   The content of a sentence is completely covered by the second sentence, but also contains additional information

4. **Meningarnas innehåll matchar helt, möjligtvis med små undantag (t. ex. pronomen, datum eller nummer)**
   The content of the sentences matches completely, possibly with minor exceptions (such as pronouns, dates or numbers)

Convenience sampling was used to gather responses, and 61 participants submitted a response to the web survey.

2.1.2. Results

The results of the human evaluation are presented in Table 1, and further illustrated in Figure 1.

MA	0.51-0.60	0.61-0.70	0.71-0.80	0.81-0.90	0.91-1.0
Mean	0.363	1.282	2.451	1.989	2.522
Std.Dev.	0.646	0.918	0.774	0.796	0.652
HA					
Mean	0.344	0.300	1.464	0.645	1.539
Std.Dev.	0.624	0.504	0.848	0.874	1.314

Table 1: Results of the human evaluation of MA and HA. Good=3, Good Partial=2, Partial=1 and Bad=0.

The sentence pairs in the corpus using the MA algorithm were generally considered more similar, than the sentence pairs of the corpus aligned with the HA algorithm.
For the MA algorithm, a sentence threshold over 0.71 seemed to produce similar sentences. The HA algorithm did not reach an average value above 2.
The high standard deviation through all intervals shows that these results should be interpreted with caution.

2.2. Evaluation II: Gold Standard

The gold standard evaluation was performed to find the best parameter settings regarding word and sentence thresholds for all three alignment algorithms (AA, MA, HA).

2.2.1. Procedure

All alignment algorithms used a threshold for word alignment and a threshold for sentence alignment. We used a gold standard to reveal the optimal combination of parameters that maximise the F1 score.
The gold standard was collected broadly following the procedure in Hwang et al. (2015), annotated by one graduate student and two payed undergraduate students. Document

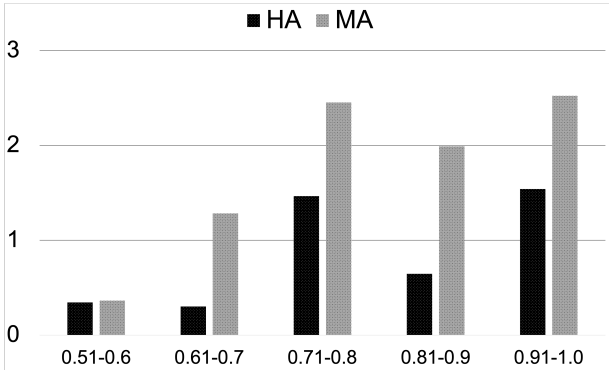

Figure 1: Average grade per interval, according to the web survey (where a value of 0 means that the sentences are not considered similar, and a value of 3 means that the sentences are considered very similar).

pairs (based on a title match) were presented to the annotators, and they were instructed to rate each sentence pair according to the descriptions of each point of the scale. If there were any doubts, they were instructed to focus on the semantic meaning rather than specific words. A training example was given prior to the annotation.
Only sentences with exactly three annotations were considered, which resulted in 4548 sentence pairs. Of these pairs, 4457 were rated as *Bad*, 37 were rated as *Bad Partial*, 24 were rated as *Good Partial*, and 30 were rated as *Good*.
The inter-annotator agreement was calculated using the Intra-class Correlation Coefficient (ICC), and revealed excellent agreement, $ICC(2,3) = 0.964$.
Since the gold standard was divided into four categories, we performed two experiments. In the first experiment (**GGPO**), the sentences rated as *Good* and *Good Partial* were considered correct alignments, and in the second experiment (**GO**) we restricted the correct alignments to only the sentences ranked as *Good*.

2.2.2. Results

As in the previous evaluation, the AA algorithm resulted in a very low number of aligned sentences for all given conditions when tested on the gold sentences.

	Max F1	No. sentences
AA	0.034	3
MA	0.758	39
HA	0.762	49

Table 2: The best-performing algorithm conditions in the GGPO setting.

In the **GGPO setting**, presented in Table 2, the results were as follows:

- The AA algorithm maximised its performance at $F1 = 0.034$, aligning 3 sentences (no difference was observed when changing parameters or vector conditions).
- The MA algorithm maximised its performance at $F1 = 0.758$, aligning 39 sentences (Mimick vectors,

word similarity threshold of 0.39, sentence similarity threshold of 0.7).

- The HA algorithm maximised its performance at $F1 = 0.762$, aligning 49 sentences (Mimick vectors, word similarity threshold of 0.79, sentence similarity threshold of 0.7).

	Max F1	No. sentences
AA	0.060	2
MA	0.892	33
HA	0.800	38

Table 3: The best-performing algorithm conditions in the GO setting.

In the **GO setting**, presented in Table 3, we saw similar tendencies:

- The AA algorithm maximised its performance at $F1 = 0.060$, aligning 2 sentences (Mimick vectors, word similarity threshold of ≥ 0.29 and sentence similarity threshold of ≥ 0.4).

- The MA algorithm maximised its performance at $F1 = 0.892$, aligning 33 sentences (Mimick vectors, word similarity threshold of ≥ 0.39 and sentence similarity threshold of 0.8).

- The HA algorithm maximised its performance at $F1 = 0.800$, aligning 38 sentences (Mimick vectors, word similarity threshold of ≥ 0.59 and sentence similarity threshold of 0.9).

Generally, the conditions using Mimick for generating vectors for out-of-vocabulary words performed better in terms of precision, recall and number of aligned sentences. The best-performing algorithm was the MA in the GO setting, and HA in the GGPO setting.

2.2.3. The Corpus

After discovering the best-performing similarity thresholds for word and sentence alignment, the winning algorithm was re-run on the raw corpus of Swedish public authorities and municipalities web texts. The performance of MA and HA did not differ much in the GGPO setting, but MA was substantially better in the GO setting. Another benefit of MA is that it less computationally demanding, which could be important to consider when running on large corpora.
We chose to run the alignment with the MA algorithm, using a word similarity threshold of 0.39 and a sentence similarity threshold of 0.7.
This resulted in a resource of 45,671 sentence pairs. After removing duplicates, 15,433 sentence pairs remained.

2.3. Evaluation III: Text Characteristics

The aligned corpus was further analysed based on text characteristics. In this evaluation, we were interested in whether the sentence pairs in the aligned resource in fact differed in complexity.

2.3.1. Procedure

Since the aligned corpus contained duplicate sentences, we only considered the 15,433 unique sentence pairs for this analysis.
First, we performed a corpus-level surface analysis, using frequency and ratio measures to get a general overview of the corpus. The corpus-level measures have been previously used for analysing comparable corpora of texts in simple and standard Swedish (Heimann Mühlenbock, 2013). However, since this corpus does not include documents, but rather sentences, some of the measures used by Heimann Mühlenbock (2013) are not applicable. The measures we excluded from the analysis were LIX (Björnsson, 1968), type-token ratio and OVIX (Hultman and Westman, 1977).
The measures used for the corpus-level analysis were:

- **Total number of words**, calculated as the number of all the alphanumeric word tokens in the sub-corpus.

- **Number of unique words**, calculated as the number of all unique alphanumeric word tokens in the sub-corpus.

- **Ratio of long words**, defined as the ratio of words longer than 6 characters to the total number of words in the sub-corpus.

- **Ratio of extra long words**, defined as the ratio of words longer than 13 characters to the total number of words in the sub-corpus.

We then performed a sentence-level surface analysis of the collected corpora. The complexity measures were calculated for all sentences in the simple Swedish sub-corpus, and all sentences in the standard sub-corpus, and significance testing was performed using two-tailed *t*-test.
The measures considered for the sentence-level surface analysis were:

- **Word length (chars)**, calculated as the mean word length in number of characters. This value was calculated for each sentence, and then averaged over the entire sub-corpus.

- **Word length (syll)**, calculated as the mean word length in number of syllables. For simplicity, we let the number of vowels correspond to the number of syllables. This value was calculated for each sentence, and then averaged over the entire sub-corpus.

- **Sentence length (words)**, calculated as the number of tokens of a sentence. This value was calculated for each sentence, and then averaged over the entire sub-corpus.

- **Number of long words**, defined as the number of words longer than 6 characters. This value was calculated for each sentence, and then averaged over the entire sub-corpus.

- **Number of extra long words**, defined as the number of words longer than 13 characters. This value was calculated for each sentence, and then averaged over the entire sub-corpus.

Finally, we calculated the measures of a subset of a feature set used for text complexity classification (Falkenjack et al., 2013). The subset (hereafter: *SCREAM-sent*) consisted of the measures that were suitable for sentence-level analysis. The selection was done according to Falkenjack (2018).
A new version of SAPIS (Fahlborg and Rennes, 2016), an API service for text analysis and simplification, was used to calculate the linguistic measures used for the *SCREAM-sent* analysis. The new version has the same functionality as the original version of SAPIS, but now uses efselab[3] (Östling, 2018) for part-of-speech tagging. SAPIS uses MaltParser (Nivre et al., 2007) version 1.9.0 for dependency parsing.
Since the *SCREAM-sent* measures were calculated at the sentence level, all measures indicating an average should be regarded as absolute for a given sentence. The significance testing was performed using two-tailed *t*-tests, assuming non-equal variances.
The selected features were:

- **avg_dep_distance_dependent**, calculated as the average dependency distance in the document.
- **avg_n_syllables**, calculated as the average number of syllables per word in the document.
- **avg_prep_comp**, calculated as the average number of prepositional complements in the document.
- **avg_sentence_depth**, calculated as the average sentence depth.
- **avg_word_length**, calculated as the average word length in a document.
- **n_content_words**, calculated as the number of content words (nouns, verbs, adjectives and adverbs).
- **n_dependencies**, calculcated as the number of dependencies.
- **n_lix_long_words**, calculated as the number of long words as defined by the LIX formula; words with more than 6 characters.
- **n_nominal_postmodifiers**, calculated as the number of nominal pre-modifiers.
- **n_nominal_premodifiers**, calculated as the number of nominal post-modifiers.
- **n_right_dependencies**, calculated as the number of right dependencies.
- **n_sub_clauses**, calculated as the number of sub-clauses.
- **Lemma frequencies**, derived from the basic Swedish vocabulary SweVoc (Heimann Mühlenbock and Johansson Kokkinakis, 2012):
  - **n_swevoc_c**, calculated as the number of words that belong to the SweVoc C word list. SweVoc C contains lemmas that are fundamental for communication.
  - **n_swevoc_d**, calculated as the number of words that belong to the SweVoc D word list. SweVoc D contains lemmas for everyday use.
  - **n_swevoc_h**, calculated as the number of words that belong to the SweVoc H word list. SweVoc H contains other highly frequent lemmas.
  - **n_swevoc_s**, calculated as the number of words that belong to the SweVoc S word list. SweVoc S contains supplementary words from Swedish Base Vocabulary Pool.
  - **n_swevoc_total**, calculated as the number of words that belong to the total SweVoc word list. SweVoc Total contains SweVoc words of all categories.
- **n_syllables**, calculated as the number of syllables in the document.
- **n_tokens**, calculated as the number of tokens in the document.
- **n_unique_tokens**, calculated as the number of unique tokens in the document.
- **n_verbal_roots**, calculated as the number of sentences where the root is a verb.
- **n_verbs**, calculated as the number of verbs.
- **right_dependency_ratio**, calculated as the ratio of the number of right dependencies to the number of total dependencies.
- **sub_clause_ratio**, calculated as the ratio of sub-clauses to the total amount of sub-clauses.
- **total_token_length**, calculated as the length of all tokens of a document.

2.3.2. Results

We performed three sets of analyses: one corpus-level surface analysis, and two sentence-level analyses. The corpus-level analysis and the first sentence-level analysis account for the measures previously used by Heimann Mühlenbock (2013). The second sentence-level analysis accounts for the *SCREAM-sent* measures.
The results of the corpus-level surface analysis are presented in Table 4. The corpus of simple sentences is slightly smaller in size regarding the total number of words. The corpus of standard sentences exhibits a larger variety regarding word variation (number of unique word tokens), and has a slightly higher ratio of long and extra long word tokens.
The results of the sentence-level surface analysis is presented in Table 5. This analysis also shows a tendency of the corpus of simple sentences to have shorter word length (in both number of characters and number of syllables),

[3] https://github.com/robertostling/efselab

Measure	*simple*	*standard*
Total number of words	177,011	181,111
Number of unique words	10,373	11,593
Ratio of long words	22.55%	22.97%
Ratio of extra long words	3.28%	3.44%

Table 4: Overview of the characteristics of the sentences in the simple part of the corpus (*simple*) and the standard part of the corpus (*standard*).

Measure	$\overline{X}_{simple}$	$\overline{X}_{standard}$	*t*	*p*
Word length (chars)	5.36	5.40	-3.03	*
Word length (syll)	1.93	1.95	-3.67	*
Sentence length (words)	11.47	11.74	-3.96	**
Number of long words	2.96	3.10	-5.66	**
Number of extra long words	0.38	0.40	-3.47	**

** p<0.05, ** p<0.001*

Table 5: Sentence-level surface analysis.

shorter sentence length and a lower number of long and extra long words. The differences are statistically significant. The results of the sentence-level analysis using the *SCREAM-sent* measures are presented in Table 6. Statistically significant *p*-values are marked in bold.

Most measures show statistically significant differences. Measures related to the length of the sentence, such as the number of syllables and the number of tokens, are generally higher in the *standard* sentences. There is also a significant difference in sentence depth and number of right dependencies, which could indicate higher complexity in the *standard* sentences. The *simple* sentences generally exhibit shorter token length, and fewer long words (>6 characters). No difference could be observed regarding the SweVoc measures from category C (core vocabulary), D (words referring to everyday objects and actions, and H (highly frequent words). However, statistically significant differences were observed for the SweVoc category S (supplementary words from the Swedish Base Vocabulary Pool), and SweVoc Total.

3. Discussion

We have presented results from three evaluations. The first and second evaluation were done on the previously aligned corpus in order to find the optimal combination of settings for the corpus alignment. Then, the corpus was aligned with the best-performing parameter settings, and the third evaluation was conducted on the new resource of aligned sentences.

- Evaluation I, the human evaluation, indicated that sentence pairs produced by the MA algorithm were regarded more similar than sentence pairs produced by the HA algorithm. A sentence similarity threshold of 0.71 seemed to produce sentence pairs that were perceived as similar, but the results lack statistical power.
- Evaluation II, the evaluation on the gold standard, indicated that the best-performing combination of settings for the alignment in the GGPO condition was the HA algorithm, using Mimick vector generation, a word similarity threshold of 0.79, and a sentence similarity threshold of 0.7. In the GO condition, the best-performing combination of settings was the MA algorithm, using Mimick vector generation, a word similarity threshold of $\geq$ 0.39 and a sentence similarity threshold of 0.8.
- Evaluation III, the evaluation of text characteristics, revealed that there are many statistically significant differences between the sentences in the simple sub-corpus and the sentences in the standard sub-corpus. The standard part of the corpus generally scores higher on features used to predict text complexity, when compared sentence-wise to the sentences collected from the material in simple Swedish.

This work has resulted in a sentence-aligned Swedish corpus of sentence pairs that differ in complexity.

Many of the differences observed in the final text complexity evaluation are to be expected if we accept the hypothesis that the sentences belonging to the standard part of the corpus are more complex than the sentences in the simple Swedish sub-corpus. Such measures include the number of long words (in characters and syllables), sentence length (in tokens and syllables), and sentence depth. However, some of the measures are not straightforward to interpret. For example, Falkenjack et al. (2013) discuss the ratio of content words to be ambiguous, since a high ratio could be indicative of higher information density, while a low ratio could mean higher syntactic complexity.

We did not observe any statistically significant differences in the majority of the SweVoc measures, and this could possibly be explained by the nature of the used alignment algorithm. Since the algorithm aims to find semantically similar sentence pairs, it is likely that the aligned sentences will also be lexically similar.

The linguistic analysis of the different parts of the corpus in this study does not include pairwise comparison, which could reveal whether the complexity differs between the sentences in the sentence pairs.

The human evaluation performed shows tendencies of when the sentences are perceived as similar. However, due to the low sample size, these tendencies can not be confirmed without an additional study with a larger sample. It would also be interesting to see whether human readers experience differences in complexity when presented with the sentences in the sentence pairs.

The collected corpus contains texts written by expert writers, following general guidelines on how to write simple text. However, even though there are some general traits of what makes a text easy to read, one must remember that the needs of the different target groups may vary. Second language learners face other problems than persons with dyslexia or aphasia, and there can be large variations within each target group. The corpus collected in this study is restricted in this sense, and future work would benefit from a more target-centred approach.

For the purpose of ATS, sentence aligned resources can be sub-optimal, since simplification operations are not limited to the sentence level. The division of long or complex sentences into multiple shorter sentences is not an uncommon

Measure	$\overline{X}_{simple}$	$\overline{X}_{standard}$	t	p
avg_dep_distance_dependent	2.44	2.46	-3.81	**
avg_n_syllables	1.80	1.81	-3.24	**
avg_prep_comp	1.46	1.51	-3.77	**
avg_sentence_depth	5.95	6.01	-2.63	*
avg_word_length	5.07	5.11	-2.91	*
n_content_words	6.64	6.77	-3.51	**
n_dependencies	13.26	13.58	-4.46	**
n_lix_long_words	2.41	2.56	-6.64	**
n_nominal_postmodifiers	0.85	0.90	-4.06	**
n_nominal_premodifiers	0.28	0.30	-3.48	**
n_right_dependencies	9.18	9.39	-4.19	**
n_sub_clauses	0.26	0.26	-0.78	
n_swevoc_c	5.38	5.46	-1.89	
n_swevoc_d	0.26	0.26	-0.02	
n_swevoc_h	0.79	0.80	-0.73	
n_swevoc_s	0.61	0.63	-2.28	*
n_swevoc_total	6.32	6.44	-2.40	*
n_syllables	21.02	21.73	-5.96	**
n_tokens	13.26	13.58	-4.46	**
n_unique_tokens	12.45	12.73	-4.64	**
n_verbal_roots	0.81	0.80	3.32	**
n_verbs	2.45	2.46	-0.45	
right_dependency_ratio	0.70	0.70	0.63	
sub_clause_ratio	0.25	0.26	-0.89	
total_token_length	62.80	65.00	-6.18	**

** p<0.05, ** p<0.001*

Table 6: Results from the t-test comparing the sentences in the simple sub-corpus (*simple*) with the sentences in the standard sub-corpus (*standard*). The *n_lix_long_words* differs from the *Number of long words* in Table 5, since the former uses the lemma form in its calculation.

operation when simplifying text, as well as the addition of explanatory sentences to clarify one complex sentence. However, it has been pointed out that certain simplification approaches are best modelled with 1-to-1 alignments (see for example Alva-Manchego et al. (2017)), and that more complex operations might need other methods and data organised in a different manner.

A resource aligned at the sentence level can be used to investigate specific sentence-level simplification operations, but it is important to be aware of the limitations, and that additional resources, such as aligned text fragments or even full documents, are needed for a complete ATS analysis.

4. Conclusion

In this article, we have presented the work on creating and evaluating an aligned resource of Swedish sentence pairs that differ in complexity. The first two evaluations aimed to find the algorithm and the best combination of parameters to maximise alignment performance. The last evaluation investigated whether the sentences in the aligned sentence pairs in fact differed in complexity.

The resulting corpus consisted of 45,671 sentence pairs, of which 15,433 were unique. The statistical analysis indicates that the sentences belonging to the simple Swedish sub-corpus are generally less complex than the sentence belonging to the standard part of the corpus, according to both surface-level measures and analysis at a deeper linguistic level.

Future research includes further analysis of the sentence pairs to see what simplification operations that are present in the data, as well as making use of this resource in data-driven text simplification research for Swedish.

5. Bibliographical References

Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C., and Specia, L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295—-305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alva-Manchego, F., Scarton, C., and Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, pages 1–87, 01.

Björnsson, C. H. (1968). *Läsbarhet*. Liber, Stockholm.

Coster, W. and Kauchak, D. (2011a). Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*, pages 1–9. Association for Computational Linguistics.

Coster, W. and Kauchak, D. (2011b). Simple english

wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Fahlborg, D. and Rennes, E. (2016). Introducing SAPIS - an API service for text analysis and simplification. In *The second national Swe-Clarin workshop: Research collaborations for the digital age, Umeå, Sweden*.

Falkenjack, J., Heimann Mühlenbock, K., and Jönsson, A. (2013). Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, number 085 in NEALT Proceedings Series 16, pages 27–40. Linköping University Electronic Press.

Falkenjack, J. (2018). Personal communication.

Fallgren, P., Segeblad, J., and Kuhlmann, M. (2016). Towards a standard dataset of swedish word vectors. In *Proceedings of the Sixth Swedish Language Technology Conference (SLTC), Umeå, Sweden*.

Heimann Mühlenbock, K. and Johansson Kokkinakis, S. (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.

Heimann Mühlenbock, K. (2013). *I see what you mean. Assessing readability for specific target groups*. Dissertation, Språkbanken, Dept of Swedish, University of Gothenburg.

Hultman, T. G. and Westman, M. (1977). *Gymnasistsvenska*. LiberLäromedel, Lund.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard wikipedia to simple wikipedia. In *HLT-NAACL*, pages 211–217.

Kajiwara, T. and Komachi, M. (2016). Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING, Osaka, Japan*, pages 1147–1158.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Pinter, Y., Guthrie, R., and Eisenstein, J. (2017). Mimicking word embeddings using subword rnns. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112.

Rennes, E. and Jönsson, A. (2016). Towards a corpus of easy to read authority web texts. In *Proceedings of the Sixth Swedish Language Technology Conference (SLTC2016), Umeå, Sweden*.

Saggion, H. (2017). *Automatic Text Simplification.* Number Vol. 32 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Scarton, C., Paetzold, G., and Specia, L. (2018). Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Song, Y. and Roth, D. (2015). Unsupervised sparse vector densification for short text similarity. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1275–1280.

Specia, L. (2010). Translating from Complex to Simplified Sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 30–39.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhang, Y., Ye, Z., Feng, Y., Zhao, D., and Yan, R. (2017). A constrained sequence-to-sequence neural model for sentence simplification. abs/1704.02312.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.

Östling, R. (2018). Part of speech tagging: Shallow or deep learning? *North European Journal of Language Technology*, 5:1–15.