Text Simplification for Language Learners: A Corpus Analysis

Sarah E. Petersen, Mari Ostendorf

Dept. of Computer Science, Dept. of Electrical Engineering
University of Washington, Seattle, WA 98195, USA
sarahs@cs.washington.edu, mo@ee.washington.edu

Abstract

Simplified texts are commonly used by teachers and students in bilingual education and other language-learning contexts. These texts are usually manually adapted, and teachers say this is a time-consuming and sometimes challenging task. Our goal is the development of tools to aid teachers by automatically proposing ways to simplify texts. As a first step, this paper presents a detailed analysis of a corpus of news articles and abridged versions written by a literacy organization in order to learn what kinds of changes people make when simplifying texts for language learners.

1. Introduction

Text simplification for second language learners is an important but somewhat controversial issue in the education community. Some experts prefer the use of “authentic” texts, i.e., texts which are written for native speakers for a purpose other than instruction, while simplified texts are also very common in educational use. Proponents of authentic texts tout positive effects on student interest and motivation and advantages of exposing students to “real” language and culture. However, authentic materials are often too hard for students who read at lower levels, as they may contain more complex language structures and vocabulary than texts intended for learners [8, 13]. Studies of the lexical and syntactic differences between authentic and simplified texts, and their effects on student comprehension, show mixed results [14, 6], suggesting that both types of texts can have a place in education. Further, authentic texts are not always available for students whose reading level does not match their intellectual level and interests. Since teachers report spending substantial amounts of time adapting texts by hand, automatic simplification could be a useful tool to help teachers adapt texts for these students.

This paper presents an analysis of a corpus of original and manually simplified news articles with the goal of gaining insight into what people most often do to simplify text in order to develop better automatic tools. When creating simplified or abridged texts, authors may drop sentences or phrases, split long sentences into multiple sentences, modify vocabulary, shorten long descriptive phrases, etc. In this paper we do not address changes to vocabulary; Burstein *et al.*’s approach to choosing synonyms for challenging words could be used to simplify vocabulary items [2]. Instead, we focus on the following research questions:

- What differences in part-of-speech usage and phrase types are found in original and simplified sentences?
- What are the characteristics of sentences which are split when an article is simplified?
- What are the characteristics of sentences which are dropped when an article is simplified?

In the following sections, we summarize prior work on text simplification, describe a sentence-aligned corpus of original and abridged articles, and provide an analysis of the article differences.

2. Related Work

Researchers have developed some text simplification systems for educational purposes. These systems rely on handwritten transformation rules. Inui *et al.* address the needs of deaf learners of written English and Japanese by paraphrasing texts to remove syntactic structures known to be difficult for this group of learners [7]. The Practical Simplification of English Text (PSET) project’s goal is to paraphrase newspaper texts for people with aphasia [3]. Max and colleagues at LIMSI-CNRS in France also target authors of texts for language-impaired readers with an interactive text simplification system that is built into a word processor to suggest simplifications while allowing the writer to maintain control over content and meaning [10]. The Educational Testing Service (ETS) has developed the Automated Text Adaptation (ATA) Tool [2], which does not directly simplify the original text but instead provides support for reading via text adaptations in English and/or Spanish that are displayed together with the original text. The adaptations include vocabulary support, marginal notes, and text-to-speech.

Most simplified texts are shorter than the original text, so extractive summarization, which selects a subset of sentences to form a summary, is a potential step in the simplification process. However, such techniques alone are insufficient for simplification, since a summarizer could choose complex sentences, resulting in a shorter text but one that is still too challenging. As our data will show, in simplification, the compression rate in words is greater than what one would obtain by simply dropping sentences.

Other research efforts are aimed at simplifying text to improve automatic language processing (vs. for the benefit of human readers). Chandrasekar and Srinivas simplify sentences with the goal of improving the performance of parsing and machine translation [4]. Their system learns transformation rules from an annotated and aligned corpus of complex sentences with manual simplifications. Siddharthan’s work on syntactic simplification and text cohesion has the goal of making individual sentences simpler without shortening the entire text. This approach uses transformations based on handwritten rules, paying particular attention to discourse-level issues in the regeneration stage (i.e., interaction between sentences in a text, not just individual sentences) [12].

We are interested in a data-driven approach to simplification like that of Chandrasekar and Srinivas. However, unlike their work, which is based on a corpus of original and manually simplified *sentences*, we study a corpus of paired *articles* in which each original sentence does not necessarily have a corresponding

Table 1: Total sentences and words and average sentence length for corpus of 104 pairs of original and abridged articles.

	Original	Abridged	Reduction
Total sentences	2539	2459	3%
Total words	41982	29584	30%
Avg. sentence length (words)	16.5	12.0	27%

Table 2: Average frequency of selected part-of-speech tags in original and abridged sentences. On average, abridged sentences contain 27% fewer words than original sentences.

Tag	Original	Abridged	Difference
Adjective	1.2	0.8	33%
Adverb	1.0	0.6	40%
CC	0.5	0.3	40%
Noun	3.6	2.8	22%
Pronoun	1.2	0.8	33%

simplified sentence. Our corpus makes it possible to learn where rewriters have dropped as well as simplified sentences. These findings should help with the development of simplification tools that can be trained from text in a variety of domains.

3. Aligned Corpus of News Articles

This work is based on a corpus of 104 original news articles with corresponding abridged versions developed by Literacyworks as part of an literacy website for learners and instructors.[1] The target audience for these articles is adult literacy learners (i.e., native speakers with poor reading skills), but the site creators suggest that the abridged articles can be used by instructors and learners of all ages. In this paper, we will refer to “original” and “abridged” texts, the terms used by the developers of this corpus. This section presents distributional characteristics of the corpus and the methods for aligning original to abridged sentences and subsentences.

3.1. Overall Corpus Statistics

Table 1 lists the total number of sentences and words and the average sentence length in words for the original and abridged portions of the corpus. There are nearly as many abridged sentence as original sentences, but there are 30% fewer words in the set of abridged articles, and the average sentence length is 27% shorter in this set.

To explore other differences between the original and abridged sentences we used an automatic parser [5] to get parses and part-of-speech (POS) tags for all sentences. Table 2 shows, for selected POS tags, the average number of that tag in the original and abridged sentences, and the percent difference between the two.[2] Since abridged sentences are on average 27% shorter, we expect fewer words and therefore fewer POS tags per sentence. However, we note that the percentage decrease in average frequency is greater for adjectives, adverbs and coordinating conjunctions, i.e., abridged sentences have fewer of these words. The percent decrease in nouns is only 22%, compared with 33% for pronouns, indicating that nouns are deleted less often than the average and it is unlikely that nouns are often replaced with pronouns. The frequency difference for determiners, IN (prepositions and subordinating conjunctions), proper nouns, and verbs is near 27%, indicating no difference other than that expected for the shorter abridged sentences. Table 3 shows average frequencies and average lengths of selected types of phrases. There are fewer phrases per sentence in the abridged sentences, and the phrases are shorter.

Table 3: Average frequency and length in words of selected phrases in original and abridged sentences.

Phrase tag	Avg phrases per sentence		Avg words per phrase	
	Original	Abridged	Original	Abridged
S	2.6	2.0	13.7	10.8
SBAR	0.8	0.5	11.3	8.5
NP	6.3	4.5	3.4	2.8
VP	3.6	2.8	9.3	7.2
PP	1.7	1.2	5.3	4.3

3.2. Alignment Methodology

In order to understand the techniques the authors used when editing each original article to create the abridged article, we hand-align the sentences in each pair of articles. This alignment was done by a native English speaker using the instructions used by Barzilay and Elhadad in their work on alignment of comparable corpora [1]. These instructions direct the annotator to mark sentences in the original and abridged versions which convey the same information in at least one clause. Since nearly all abridged sentences have a corresponding original sentence but some original sentences are dropped, we asked the annotator to align all the sentences in each abridged article to a corresponding sentence or sentences in the original file and automatically reversed the hand alignments to get the alignments of original to abridged sentences.

3.3. Original and Aligned Sentences

Table 4 shows the distribution of original sentences into categories based on the alignment described above.[3] Sentences can be dropped (no corresponding abridged sentence) or aligned to one or more abridged sentences. For sentences which are aligned to exactly one other sentence, we calculate whether the abridged sentence is more than 20% shorter or longer, or approximately the same length as the original sentence. A sentence which is aligned to more than one abridged sentence is hypothesized to be split. Likewise, sentences which are aligned to a single shorter sentence are hypothesized to be split with one part dropped. Note that the average sentence length in these categories is longer than the other categories. We will further investigate the alignment of these hypothesized split sentences in the next subsection. The last two categories in the table are rare and specific to the abilities of human authors. An original sentence aligned to a longer abridged sentence indicates that the author added some material, perhaps to explain a difficult point. The average length of these original sentences is shorter than the other categories. In the case of merged sentences,

[1] http://literacynet.org/cnnsf/index_cnnsf.html (2004-2007)

[2] We use the standard Penn Treebank tag set, aggregating tags for adjectives, adverbs, determiners, nouns, proper nouns, pronouns and verbs. For example, the adjective category includes JJ, JJR and JJS.

[3] Since multiple sentences in an article could be considered to contain the same information per the alignment instructions, there are infrequent overlaps between categories which cause the percentages to add up to slightly more than 100%, e.g., a sentence could be both split and merged.

Table 4: Number of cases and average sentence length in words for different categories of alignment of original to abridged sentences.

Category	Num Sentences	Avg Length
Total	2539 (100%)	16.5
1 to 0 (dropped)	763 (30%)	14.1
1 to ≥2 (split)	470 (19%)	24.6
1 to 1 (total)	1188 (47%)	15.8
1 to 1 (shorter abr.)	350 (14%)	21.0
1 to 1 (same length abr.)	725 (29%)	14.4
1 to 1 (longer abr.)	113 (4%)	9.1
2 to 1 (merged)	167 (7%)	14.6

Table 5: Examples of "split", "edited", and "different" aligned sentences. The split point in the first sentence is indicated by ***.

Split	
Original	Keith Johnson is the Makah Tribe Spokesman, *** and he comments, "We made history today.
Abridged	Keith Johnson is the Makah Tribe Spokesman. He said, "We made history today.
Edited	
Original	Congress gave Yosemite the money to repair damage from the 1997 flood.
Abridged	Congress gave the money after the 1997 Flood.
Different	
Original	The park service says the solution is money.
Abridged	Why hasn't the National Park Service kept up the park repairs? There is a lack of money.

two sentences are aligned to one abridged sentence. These cases are much more difficult to handle, but relatively infrequent, so we will focus the remainder of our analysis on the other categories.

3.4. Annotating True Split Sentences

Nearly 20% of the original sentences are aligned to more than one abridged sentence. We hypothesize that many of these cases are long sentences with multiple clauses that the author chose to split into shorter sentences. Similarly, sentences that are aligned to much shorter abridged sentences are likely to be split with one part dropped. To test these assumptions, we asked the annotator to mark split points in the sentences corresponding to their alignment to the abridged sentences. Some sentences did not have split points and were categorized as either "edited," indicating that the sentence has minor changes but no obvious split point, or "different," indicating that the sentences convey the same information though the wording is very different. Table 5 shows examples of these three categories. Table 6 shows the distribution of original sentences in each category for the hypothesized one-to-many and one-to-one splits. In the one-to-one case, all true splits consist of a piece that is kept and a piece that is dropped. The vast majority of the one-to-many case and the majority of the one-to-one case are true splits, so finding sentences to split is clearly an important task. Potential clues to split sentences are explored in the next section.

Table 6: Distribution of hypothesized split sentences.

Category	Num Sentences	
	One to Many	One to One
Total	470 (100%)	350 (100%)
True split	368 (78%)	202 (58%)
Edited	17 (4%)	145 (41%)
Different	85 (18%)	3 (1%)

4. Analysis of Split vs. Unsplit Sentences

A step in automatic simplification is choosing sentences to split. We expect that long sentences would be selected for splitting, but other characteristics are likely to be considered, too. In this section we analyze the 570 sentences identified as "true splits" compared to 1205 unsplit sentences. For the purpose of this analysis, the "dropped" sentences are not included in the unsplit category, since some of them might have characteristics of sentences that should be split if they were kept. Other sentence categories from Table 4 and the edited and different sentences from the hypothesized split sentences are considered unsplit. As expected, split sentences are longer, with an average sentence length of 23.3 words compared to 15.0 words for unsplit sentences. The average number of phrases identified by the parser (S, NP, etc.) and the length of these phrases is also longer on average for split sentences.

In addition to length, we hypothesize that the decision to split a sentence is based on syntactic features, since splitting a sentence reduces the syntactic complexity of the resulting sentences while retaining the same information. To investigate which features are most important for splitting sentences, we used the C4.5 decision tree learner[4] to build a classifier for split and unsplit sentences. We chose C4.5 and its rule generator because the results are easily interpreted, and the focus of this work is on analysis rather than classification. We use the following features for each sentence:

- Sentence length in words.
- Number of: adjectives, adverbs, CC, IN, determiners, nouns, proper nouns, pronouns, and verbs.
- Number and average length of: S, SBAR, NP, PP, and VP.

The POS tags are aggregated as described in Section 3.1 and the phrase types are chosen because they show differences in average frequency and length in the original and abridged sentences.

Using a pruning level chosen by 10-fold cross-validation (CV), we trained a tree and its accompanying rules on the entire split vs. unsplit dataset. Average CV error rate is 29%; these classes are not easily separable. As expected, length is the most important feature, leading to the two rules with highest usage: sentences with length less than 19 tend not to be split, and sentences with length more than 24 tend to be split. On the training set, these rules apply to 907 and 329 sentences, respectively, with error rates of 17% and 33%. The next most frequently applied rule for unsplit sentences uses features for length < 24 and NP average length $<= 1.4$, i.e., moderately long sentences with short noun phrases, suggesting that longer noun phrases tend to lead to a split. Commonly used features in other rules for split sentences include the number of nouns, pronouns, verbs, determiners and VPs. Surprisingly, S and SBAR were not commonly used features.

These observations give us insight into characteristics of sentences that are split, but choosing such sentences is only the first

[4] http://www.rulequest.com/Personal/

step. It is also necessary to transform the resulting fragments into complete, grammatical sentences. We tested Siddharthan's syntactic simplifier [12], which does both of these steps, on our corpus and found that it split a comparable number of sentences as the authors. Hand-analysis of the split sentences in 5 articles showed that 88% were "good splits" and that the average grammar score on a 0-2 scale (0 for bad, 2 for good, and 1 for imperfect but understandable) was 1.3. Similar results were found when the system was run on other news and science texts.[5]

5. Analysis of Dropped Sentences

Most abridged or simplified texts are shorter than the original text, so another step to consider is choosing sentences to omit. In this section, we compare the dropped sentences with the rest of the original sentences. As in the previous section, we use the C4.5 rule generator to see which features are most important.

In the case of dropped sentences, the rationale for picking sentences to drop is more likely to be content-based than syntactic, thus we consider a different set of features in this section. Intuitively, we expect that sentences which are somewhat redundant may be dropped. However, it is also possible that some repetition is a good thing since familiar content may be easier to read. To explore these options, we use features for the percentage of the content words in a sentence that have already occurred in previous sentences in the document. Position in the original document may also be a factor. The complete set of features for each sentence is:

- Position in the document: sentence number, percentage.
- Paragraph number, first or last sentence in paragraph.
- Does this sentence contain a direct quotation?
- Percentage of words that are stop words (Wordnet 1.6 list).
- Percentage of content words which have already occurred one/two/three/four/more times in the document.

Again, we choose a pruning threshold using CV and train a classifier on the entire dataset. The classifier performance is little better than always choosing the majority class (not dropped). The rule with highest applicability (895 sentences) and lowest error rate (15%)

> position $\leq$ 12, stop words $\leq$ 70%, content words seen once $\leq$ 40%, no content words seen $>$ 5 times

indicates that sentences early in the document with low redundancy are not dropped. Rules for dropped sentences have lower applicability and higher error rates. The quote feature is used several times; we have observed that quotes are often removed from the abridged documents in this corpus, and the rules indicate that this is particularly true if the quote consists of more than 70% stop words or is past the 12th sentence in the document. Another rule for dropped sentences is that later sentences (position $>$ 35) are dropped; it applies to 91 cases with 33% error. The redundancy features are also used, though with higher error rates.

We also explored using an existing extractive summarization tool, e.g., [11], applying it to five articles from the Literacyworks corpus and using the ROUGE evaluation tool [9] to compare the resulting summaries to the handwritten abridged versions (i.e., using a single reference). The average recall for bigram co-occurrence is 24%, with 50% recall for longest common subsequence. For a set of 10 articles from another corpus, with three handwritten references each, we get 48% recall for bigram co-occurrence and 63% for longest common subsequence. These are promising results for the use of summarization as one step in a simplification system.

6. Summary

We have described a corpus of original and abridged news articles, observing that the simplified sentences contain fewer adverbs and coordinating conjunctions, in particular, and fewer and shorter phrases of all types. Our analyses comparing the original and abridged articles shows the importance of syntactic features (in addition to sentence length) for decisions about sentence splitting, and of position and redundancy information in decisions about which sentences to keep and which to drop. These insights will be useful in future work on automating the simplification process, but may also be useful for improving reading level detection.

Acknowledgments: The authors thank P. Heavenridge (Literacyworks) for use of the corpus, and B. Blaser for the hand alignments. This work was supported by the National Science Foundation (NSF) under Grant No. IIS-0326276. Any opinions or findings expressed here are those of the authors and do not necessarily reflect the views of the NSF.

7. References

[1] Barzilay, R. & Elhadad, N. Sentence alignment for monolingual comparable corpora. *Proc. of EMNLP*, pp. 25-32, 2003.

[2] Burstein, J. *et al.* The Automated Text Adaptation Tool. *Demo Proc. NAACL-HLT*, 2007.

[3] Canning, Y. *et al.* Cohesive regeneration of syntactically simplified newspaper text. *Proc. ROMAND*, pp. 3-16, 2000.

[4] Chandrasekar, R. & Srinivas, B. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10:183-190, 1997.

[5] Charniak, E. A maximum-entropy-inspired parser. *Proc. NAACL*, pp. 132-139, 2000.

[6] Crossley, S. *et al.* A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, 91(1):15-30, 2007.

[7] Inui, K. *et al.* Text simplification for reading assistance: A project note. *Proc. IWP2003*, pp. 9-16, 2003.

[8] Kilickaya, F. Authentic materials and cultural content in EFL classrooms. *The Internet TESL Journal*, X(7), 2004.

[9] Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. *Proc. ACL Workshop: Text Summarization Branches Out*, 2004.

[10] Max, A. Writing for language-impaired readers. *Proc. CICLing*, vol. LNCS 3878.

[11] Marcu, D. Improving summarization through rhetorical parsing tuning. *The Sixth Workshop on Very Large Corpora*, pp. 206-215, 1998.

[12] Siddharthan, A. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77-109, 2006.

[13] Sonmez, S. An overview of studies on the use of authentic texts in language classrooms. *Proc. Online Conf. Second & Foreign Language Teaching & Research*, pp. 51-62, 2007.

[14] Young, D. Linguistic simplification of SL reading material: Effective instructional practice? *Modern Language Journal*, 83(3):350-366, 1999.

[5] The split judgments were done by the same annotator who did the alignment and split-point annotation for the corpus.