Problems in Current Text Simplification Research: New Data Can Help

Wei Xu[1] and **Chris Callison-Burch**[1] and **Courtney Napoles**[2]
[1] Computer and Information Science Department
University of Pennsylvania
{xwe, ccb}@seas.upenn.edu
[2] Department of Computer Science
Johns Hopkins University
courtneyn@jhu.edu

Abstract

Simple Wikipedia has dominated simplification research in the past 5 years. In this opinion paper, we argue that focusing on Wikipedia limits simplification research. We back up our arguments with corpus analysis and by highlighting statements that other researchers have made in the simplification literature. We introduce a new simplification dataset that is a significant improvement over Simple Wikipedia, and present a novel quantitative-comparative approach to study the quality of simplification data resources.

1 Introduction

The goal of text simplification is to rewrite complex text into simpler language that is easier to understand. Research into this topic has many potential practical applications. For instance, it can provide reading aids for people with disabilities (Carroll et al., 1999; Canning et al., 2000; Inui et al., 2003), low-literacy (Watanabe et al., 2009; De Belder and Moens, 2010), non-native backgrounds (Petersen and Ostendorf, 2007; Allen, 2009) or non-expert knowledge (Elhadad and Sutaria, 2007; Siddharthan and Katsos, 2010). Text simplification may also help improve the performance of many natural language processing (NLP) tasks, such as parsing (Chandrasekar et al., 1996), summarization (Siddharthan et al., 2004; Klebanov et al., 2004; Vanderwende et al., 2007; Xu and Grishman, 2009), semantic role labeling (Vickrey and Koller, 2008), information extraction (Miwa et al., 2010) and machine translation (Gerber and Hovy, 1998; Chen et al., 2012), by transforming long, complex sentences into ones that are more easily processed.

The Parallel Wikipedia Simplification (PWKP) corpus prepared by Zhu et al. (2010), has become the benchmark dataset for training and evaluating automatic text simplification systems. An associated test set of 100 sentences from Wikipedia has been used for comparing the state-of-the-art approaches. The collection of simple-complex parallel sentences sparked a major advance for machine translation-based approaches to simplification. However, we will show that this dataset is deficient and should be considered obsolete.

In this opinion paper, we argue that Wikipedia as a simplification data resource is suboptimal for several reasons: 1) It is prone to automatic sentence alignment errors; 2) It contains a large proportion of inadequate simplifications; 3) It generalizes poorly to other text genres. These problems are largely due to the fact that Simple Wikipedia is an encyclopedia spontaneously and collaboratively created for "children and adults who are learning English language" without more specific guidelines. We quantitatively illustrate the seriousness of these problems through manual inspection and statistical analysis.

Our manual inspection reveals that about 50% of the sentence pairs in the PWKP corpus are not simplifications. We also introduce a new comparative approach to simplification corpus analysis. In particular, we assemble a new simplification corpus of news articles,[1] re-written by professional editors to meet the readability standards for children at multi-

[1]This Newsela corpus can be requested following the instructions at: https://newsela.com/data/

Transactions of the Association for Computational Linguistics, vol. 3, pp. 283–297, 2015. Action Editor: Rada Mihalcea.
Submission batch: 12/2014; Revision batch 4/2015; Published 5/2015.
©2015 Association for Computational Linguistics. Distributed under a CC-BY-NC-SA 4.0 license.
Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Not Aligned (17%)		[NORM] The soprano ranges are also written from middle C to A an octave higher, but sound one octave higher than written. [SIMP] The xylophone is usually played so that the music sounds an octave higher than written.
Not Simpler (33%)		[NORM] Chile is the longest north-south country in the world, and also claims of Antarctica as part of its territory. [SIMP] Chile, which claims a part of the Antarctic continent, is the longest country on earth.
		[NORM] Death On 1 October 1988, **Strauss** collapsed while hunting with the Prince of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg. [SIMP] Death On October 1, 1988, **Strauß** collapsed while hunting with the Prince of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg.
Real Simplification (50%)	**Deletion Only (21%)**	[NORM] This **article** is a list of the 50 U.S. states **and the District of Columbia** ordered by population density. [SIMP] This is a list of the 50 U.S. states**,** ordered by population density.
	Paraphrase Only (17%)	[NORM] In 2002, both Russia and China also had **prison populations in excess of 1 million**. [SIMP] In 2002, both Russia and China also had **over 1 million people in prison**.
	Deleltion + (12%) Paraphrase	[NORM] All adult Muslims**, with exceptions for the infirm, are required to offer** Salat prayers five times **daily**. [SIMP] All adult Muslims **should do** Salat prayers five times **a day**.

Table 1: Example sentence pairs (NORM-SIMP) aligned between English Wikipedia and Simple English Wikipedia. The breakdown in percentages is obtained through manual examination of 200 randomly sampled sentence pairs in the Parallel Wikipedia Simplification (PWKP) corpus.

ple grade levels. This parallel corpus is higher quality and its size is comparable to the PWKP dataset. It helps us to showcase the limitations of Wikipedia data in comparison and it provides potential remedies that may improve simplification research.

We are not the only researchers to notice problems with Simple Wikipedia. There are many hints in past publications that reflect the inadequacy of this resource, which we piece together in this paper to support our arguments. Several different simplification datasets have been proposed (Bach et al., 2011; Woodsend and Lapata, 2011a; Coster and Kauchak, 2011; Woodsend and Lapata, 2011b), but most of these are derived from Wikipedia and not thoroughly analyzed. Siddharthan (2014)'s excellent survey of text simplification research states that one of the most important questions that needs to be addressed is "how good is the quality of Simple English Wikipedia". To the best of our knowledge, we are the first to systematically quantify the quality of Simple English Wikipedia and directly answer this question.

We make our argument not as a criticism of others or ourselves, but as an effort to refocus research directions in the future (Eisenstein, 2013). We hope to inspire the creation of higher quality simplification datasets, and to encourage researchers to think critically about existing resources and evaluation methods. We believe this will lead to breakthroughs in text simplification research.

2 Simple Wikipedia is not that simple

The Parallel Wikipedia Simplification (PWKP) corpus (Zhu et al., 2010) contains approximately 108,000 automatically aligned sentence pairs from cross-linked articles between Simple and Normal English Wikipedia. It has become a benchmark dataset for simplification largely because of its size and availability, and because follow-up papers (Woodsend and Lapata, 2011a; Coster and Kauchak, 2011; Wubben et al., 2012; Narayan and Gardent, 2014; Siddharthan and Angrosh, 2014; Angrosh et al., 2014) often compare with Zhu et al.'s system outputs to demonstrate further improvements.

The large quantity of parallel text from Wikipedia made it possible to build simplification systems using statistical machine translation (SMT) technology. But after the initial success of these first-generation systems, we started to suffer from the

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

inadequacy of the parallel Wikipedia simplification datasets. There is scattered evidence in the literature. Bach et al. (2011) mentioned they have attempted to use parallel Wikipedia data, but opted to construct their own corpus of 854 sentences (25% from New York Times and 75% are from Wikipedia) with one manual simplification per sentence. Woodsend and Lapata (2011a) showed that rewriting rules learned from Simple Wikipedia revision histories produce better output compared to the "unavoidably noisy" aligned sentences from Simple-Normal Wikipedia. The Woodsend and Lapata (2011b) model, that used quasi-synchronous grammars learned from Wikipedia revision history, left 22% sentences unchanged in the test set. Wubben et al. (2012) found that a phrase-based machine translation model trained on the PWKP dataset often left the input unchanged, since "much of training data consists of partially equal input and output strings". Coster and Kauchak (2011) constructed another parallel Wikipedia dataset using a more sophisticated sentence alignment algorithm with an additional step that first aligns paragraphs. They noticed that 27% aligned sentences are identical between simple and normal, and retained them in the dataset "since not all sentences need to be simplified and it is important for any simplification algorithm to be able to handle this case". However, we will show that many sentences that need to be simplified are not simplified in the Simple Wikipedia.

We manually examined the Parallel Wikipedia Simplification (PWKP) corpus and found that it is noisy and half of its sentence pairs are not simplifications (Table 1). We randomly sampled 200 one-to-one sentence pairs from the PWKP dataset (one-to-many sentence splitting cases consist of only 6.1% of the dataset), and classify each sentence pair into one of the three categories:

Not Aligned (17%) -
Two sentences have different meanings, or only have partial content overlap.

Not Simpler (33%)-
The SIMP sentence has the same meaning as the NORM sentence, but is not simpler.

Real Simplification (50%)-
The SIMP sentence has the same meaning as the NORM sentence, and is simpler. We further breakdown into whether the simplification is due to deletion or paraphrasing.

Table 1 shows a detailed breakdown and representative examples for each category. Although Zhu et al. (2010) and Coster and Kauchak (2011) have provided a simple analysis on the accuracy of sentence alignment, there are some important facts that cannot be revealed without in-depth manual inspection. The "non-simplification" noise in the parallel Simple-Normal Wikipedia data is a much more serious problem than we all thought. The quality of "real simplifications" also varies: some sentences are simpler by only one word while the rest of sentence is still complex.

The main causes of non-simplifications and partial-simplifications in the parallel Wikipedia corpus include: 1) The Simple Wikipedia was created by volunteer contributors with no specific objective; 2) Very rarely are the simple articles complete re-writes of the regular articles in Wikipedia (Coster and Kauchak, 2011), which makes automatic sentence alignment errors worse; 3) As an encyclopedia, Wikipedia contains many difficult sentences with complex terminology. The difficulty of sentence alignment between Normal-Simple Wikipedia is highlighted by a recent study by Hwang et al. (2015) that achieves state-of-the-art performance of 0.712 maximum F1 score (over the precision-recall curve) by combining Wiktionary-based and dependency-parse-based sentence similarities. And in fact, even the simple side of the PWKP corpus contains an extensive English vocabulary of 78,009 unique words. 6,669 of these words do not exist in the normal side (Table 2). Below is a sentence from an article entitled "Photolithography" in Simple Wikipedia:

Microphototolithography is the use of photolithography to transfer geometric shapes on a photomask to the surface of a semiconductor wafer for making integrated circuits.

We should use the PWKP corpus with caution and consider other alternative parallel simplification corpora. Alternatives could come from Wikipedia (but better aligned and selected) or from manual simplification of other domains, such as newswire. In the

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

PWKP	Normal	Simple
#words (avg. freq)	95,111 (23.91)	78,009 (23.88)
Normal	0	6,669(1.31)
Simple	23,771 (1.42)	0

Table 2: The vocabulary size of the Parallel Wikipedia Simplification (PWKP) corpus and the vocabulary difference between its normal and simple sides (as a 2×2 matrix). Only words consisting of the 26 English letters are counted.

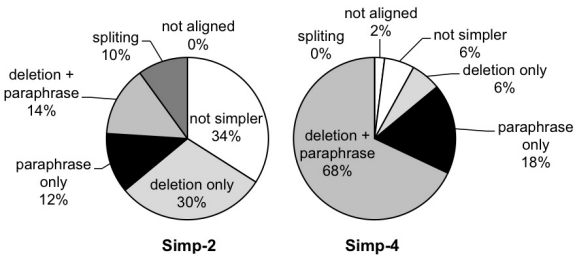

Figure 1: Manual classification of aligned sentence pairs from the Newsela corpus. We categorize randomly sampled 50 sentence pairs drawn from the Original-Simp2 and 50 sentences from the Original-Simp4.

next section, we will present a corpus of news articles simplified by professional editors, called the Newsela corpus. We perform a comparative corpus analysis of the Newsela corpus versus the PWKP corpus to further illustrate concerns about PWKP's quality.

3 What the Newsela corpus teaches us

To study how professional editors conduct text simplification, we have assembled a new simplification dataset that consists of 1,130 news articles. Each article has been re-written 4 times for children at different grade levels by editors at Newsela[2], a company that produces reading materials for pre-college classroom use. We use Simp-4 to denote the most simplified level and Simp-1 to denote the least simplified level. This data forms a parallel corpus, where we can align sentences at different reading levels, as shown in Table 3.

Unlike Simple Wikipedia, which was created without a well-defined objective, Newsela is meant to help teachers prepare curricula that match the English language skills required at each grade level. It is motivated by the Common Core Standards (Porter et al., 2011) in the United States. All the Newsela articles are grounded in the Lexile[3] readability score, which is widely used to measure text complexity and assess students' reading ability.

3.1 Manual examination of Newsela corpus

We conducted a manual examination of the Newsela data similar to the one for Wikipedia data in Table 1. The breakdown of aligned sentence pairs between different versions in Newsela is shown in Figure 1. It is based on 50 randomly selected sentence pairs and shows much more reliable simplification than the Wikipedia data.

We designed a sentence alignment algorithm for the Newsela corpus based on Jaccard similarity (Jaccard, 1912). We first align each sentence in the simpler version (e.g. $s1$ in Simp-3) to the sentence in the immediate more complex version (e.g. $s2$ in Simp-2) of the highest similarity score. We compute the similarity based on overlapping word lemmas:[4]

$$Sim(s1, s2) = \frac{|Lemmas(s1) \cap Lemmas(s2)|}{|Lemmas(s1) \cup Lemmas(s2)|} \quad (1)$$

We then align sentences into groups across all 5 versions for each article. For cases where no sentence splitting is involved, we discard any sentence pairs with a similarity smaller than 0.40. If splitting occurs, we set the similarity threshold to 0.20 instead.

Newsela's professional editors produce simplifications with noticeably higher quality than Wikipedia's simplifications. Compared to sentence alignment for Normal-Simple Wikipedia, automatically aligning Newsela is more straightforward and reliable. The better correspondence between the simplified and complex articles and the availability of multiple simplified versions in the Newsela data also contribute to the accuracy of sentence alignment.

[2] https://newsela.com/

[3] http://en.wikipedia.org/wiki/Lexile

[4] We use the WordNet lemmatization in the NLTK package: http://www.nltk.org/

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Grade Level	Lexile Score	Text
12	1400L	**Slightly more** fourth-graders **nationwide are reading proficiently compared with** a decade ago**, but** only a third of them **are now reading well**, according to a new report.
7	1070L	Fourth-graders **in most states are better readers than they were a decade** ago**. But** only a third of them **actually are able to read well**, according to a new report.
6	930L	Fourth-graders **in most states** are better readers than they were a decade ago. But **only a third of them** actually **are able to** read well**, according to a new report**.
4	720L	**Most** fourth-graders are better readers than **they were** 10 years ago. But **few of them can actually** read well.
3	510L	Fourth-graders are better readers than 10 years ago. But few of them read well.

Table 3: Example of sentences written at multiple levels of text complexity from the Newsela data set. The Lexile readability score and grade level apply to the whole article rather than individual sentences, so the same sentences may receive different scores, e.g. the above sentences for the 6th and 7th grades. The bold font highlights the parts of sentence that are different from the adjacent version(s).

	Newsela					PWKP	
	Original	Simp-1	Simp-2	Simp-3	Simp-4	Normal	Simple
Total #sents	56,037	57,940	63,419	64,035	64,162	108,016	114,924
Total #tokens	1,301,767	1,126,148	1,052,915	903,417	764,103	2,645,771	2,175,240
Avg #sents per doc	49.59	51.27	56.12	56.67	56.78	—	—
Avg #words per doc	1,152.01	996.59	931.78	799.48	676.2	—	—
Avg #words per sent	**23.23**	19.44	16.6	14.11	**11.91**	***24.49**	***18.93**
Avg #chars per word	**4.32**	4.28	4.21	4.11	**4.02**	**5.06**	**4.89**

Table 4: Basic statistics of the Newsela Simplification corpus vs. the Parallel Wikipedia Simplification (PWKP) corpus. The Newsela corpus consists of 1130 articles with original and 4 simplified versions each. Simp-1 is of the least simplified level, while Simp-4 is the most simplified. The numbers marked by * are slightly different from previously reported, because of the use of different tokenizers.

Newsela	Original	Simp-1	Simp-2	Simp-3	Simp-4
#words (avg. freq)	**39,046 (28.31)	33,272 (28.64)	29,569 (30.09)	24,468 (31.17)	20,432 (31.45)
Original	0	724 (1.19)	815 (1.25)	720 (1.32)	*583 (1.33)
Simp-1	6,498 (1.38)	0	618 (1.08)	604 (1.15)	521 (1.21)
Simp-2	10,292 (1.67)	4,321 (1.32)	0	536 (1.13)	475 (1.16)
Simp-3	15,298 (2.14)	9,408 (1.79)	5,637 (1.46)	0	533 (1.14)
Simp-4	**19,197 (2.60)	13,361 (2.24)	9,612 (1.87)	4,569 (1.40)	0

Table 5: This table shows the vocabulary changes between different levels of simplification in the Newsela corpus (as a 5×5 matrix). Each cell shows the number of unique word types that appear in the corpus listed in the column but do not appear in the corpus listed in the row. We also list the average frequency of those vocabulary items. For example, in the cell marked *, the Simp-4 version contains 583 unique words that do not appear in the Original version. By comparing the cells marked **, we see about half of the words (19,197 out of 39,046) in the Original version are not in the Simp-4 version. Most of the vocabulary that is removed consists of low-frequency words (with an average frequency of 2.6 in the Original).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

3.2 Vocabulary statistics

Table 4 shows the basic statistics of the Newsela corpus and the PWKP corpus. They are clearly different. Compared to the Newsela data, the Wikipedia corpus contains remarkably longer (more complex) words and the difference of sentence length before and after simplification is much smaller. We use the Penn Treebank tokenizer in the Moses package.[5]

Tables 2 and 5 show the vocabulary statistics and the vocabulary difference matrix of the PWKP and Newsela corpus. While the vocabulary size of the PWKP corpus drops only 18% from 95,111 unique words to 78,009, the vocabulary size of the Newsela corpus is reduced dramatically by 50.8% from 39,046 to 19,197 words at its most simplified level (Simp-4). Moreover, in the Newsela data, only several hundred words that occur in the simpler version do not occur in the more complex version. The words introduced are often abbreviations ("National Hurricane Center" $\rightarrow$ "NHC"), less formal words ("unscrupulous" $\rightarrow$ "crooked") and shortened words ("chimpanzee" $\rightarrow$ "chimp"). This implies a more complete and precise degree of simplification in the Newsela than the PWKP dataset.

3.3 Log-odds-ratio analysis of words

In this section, we visualize the differences in the topics and degree of simplification between the Simple Wikipedia and the Newsela corpus. To do this, we employ the log-odds-ratio informative Dirichlet prior method of Monroe et al. (2008) to find words and punctuation marks that are statistically overrepresented in the simplified text compared to the original text. The method measures each token by the z-score of its log-odds-ratio as:

$$\frac{\delta_t^{(i-j)}}{\sqrt{\sigma^2(\delta_t^{(i-j)})}} \tag{2}$$

It uses a background corpus when calculating the log-odds-ratio δ_t for token t, and controls for its variance σ^2. Therefore it is capable of detecting differences even in very frequent tokens. Other methods used to discover word associations, such as mutual information, log likelihood ratio, t-test and chi-square, often have problems with frequent words (Jurafsky et al., 2014). We choose the Monroe et al. (2008) method because many function words and punctuations are very frequent and play important roles in text simplification.

The log-odds-ratio $\delta_t^{(i-j)}$ for token t estimates the difference of the frequency of token t between two text sets i and j as:

$$\begin{aligned}\delta_t^{(i-j)} = &\log(\frac{y_t^i + \alpha_t}{n^i + \alpha_0 - (y_t^i + \alpha_t)}) \\ &- \log(\frac{y_t^j + \alpha_t}{n^j + \alpha_0 - (y_t^j + \alpha_t)})\end{aligned} \tag{3}$$

where n^i is the size of corpus i, n^j is the size of corpus j, y_t^i is the count of token t in corpus i, y_t^j is the count of token t in corpus j, α_0 is the size of the background corpus, and α_t is the count of token t in the background corpus. We use the combination of both simple and complex sides in the corpus as the background.

And the variance of the log-odds-ratio is estimated by:

$$\sigma^2(\delta_t^{(i-j)}) \approx \frac{1}{y_t^i + \alpha_t} + \frac{1}{y_t^j + \alpha_t} \tag{4}$$

Table 6 lists the top 50 words and punctuation marks that are the most strongly associated with the complex text. Both corpora significantly reduce function words and punctuation. The content words show the differences of the topics and subject matters between the two corpora. Table 7 lists the top 50 words that are the most strongly associated with the simplified text. The two corpora are more agreeable on what the simple words are than what complex words need to be simplified.

Table 8 shows the frequency and odds ratio of example words from the top 50 complex words. The odds ratio of token t between two texts sets i and j is defined as:

$$r_t^{(i-j)} = \frac{y_t^i / y_t^j}{n^i / n^j} \tag{5}$$

It reflects the difference of topics and degree of simplification between the Wikipedia and the Newsela data. The high proportion of clause-related function words, such as "which" and "where",

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Linguistic class	Newsela - Original	Wikipedia (PWKP) - Normal
Punctuation	, "– ; ’ ( )	, ; –
Determiner/Pronoun	**which** we an such who i that a whose	**which** whom
Contraction	’s	
Conjunction	**and while although**	**and although while**
Prepositions	**of as including with** according by among in despite	**as with** following to **of within** upon **including**
Adverb		currently approximately initially primarily subsequently typically thus formerly
Noun	percent director data research decades industry policy development state decade status university residents	film commune footballer pays-de-la-loire walloon links midfielder defender goalkeeper
Adjective	federal potential recent executive economic	northern northwestern southwestern external due numerous undated various prominent
Verb	advocates based access	referred derived established situated considered consists regarded having

Table 6: Top 50 tokens associated with the complex text, computed using the Monroe et al. (2008) method. Bold words are shared by the complex version of Newsela and the complex version of Wikipedia.

Linguistic class	Newsela - Simp4	Wikipedia (PWKP) - Simple
Punctuation	.	.
Determiner/Pronoun	**they it he she** them **lot**	**it he they lot** this **she**
Conjunction		because
Adverb	also not **there** too **about very** now then how	**about very there**
Noun	**people** money scientists government things countries rules problems group	movie **people** northwest north region loire player websites southwest movies football things
Adjective	**many important big** new used	**big** biggest famous different **important many**
Verb	**is are can** will **make get** were wants **was called** help hurt be **made like** stop want works do live	found **is made called** started pays said **was** got **are like get can** means says has went comes **make** put used

Table 7: Top 50 tokens associated with the simplified text.

	Newsela			PWKP		
	Original	Simp-4	odds-ratio	Normal	Simple	odds-ratio
which	2259	249	0.188	7261	4608	0.774
where	1472	546	0.632	1972	1470	0.909
advocates	136	0	0	6	3	0.610
approximately	21	0	0	480	140	0.356
thus	35	9	0.438	385	138	0.437

Table 8: Frequency of example words from Table 6. These complex words are reduced at a much greater rate in the simplified Newsela than they are in the Simple English Wikipedia. A smaller odds ratio indicates greater reduction.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Newsela - Original	Wikipedia (PWKP) - Normal		Newsela - Simp4	Wikipedia (PWKP) - Simple
PP(of) → IN NP	**PP(as) → IN NP**	1	**S(is) → NP VP .**	**NP(it) → PRP**
WHNP(which) → WDT	**PP(of) → IN NP**	2	**NP(they) → PRP**	**S(is) → NP VP .**
SBAR(which) → WHNP S	VP(born) → VBN NP NP PP	3	**S(are) → NP VP .**	**S(was) → NP VP .**
PP(to) → TO NP	**WHNP(which) → WDT**	4	**S(was) → NP VP .**	**NP(he) → PRP**
NP(percent) → CD NN	**PP(to) → TO NP**	5	**NP(people) → NNS**	**NP(they) → PRP**
WHNP(that) → WDT	NP(municipality) → DT JJ NN	6	VP(is) → VBZ NP	NP(player) → DT JJ JJ NN NN
SBAR(that) → WHNP S	FRAG(-) → ADJP :	7	**NP(he) → PRP**	**S(are) → NP VP .**
PP(with) → IN NP	FRAG(-) → FRAG : FRAG	8	S(were) → NP VP .	NP(movie) → DT NN
PP(according) → VBG PP	NP()) → NNP NNP NNP	9	**NP(it) → PRP**	**S(has) → NP VP .**
NP(percent) → NP PP	NP(film) → DT NN	10	**S(can) → NP VP .**	**VP(called) → VBN NP**
NP(we) → PRP	NP(footballer) → DT JJ JJ NN	11	S(will) → NP VP .	VP(is) → VBZ PP
PP(including) → VBG NP	NP(footballer) → NP SBAR	12	ADVP(also) → RB	VP(made) → VBN PP
SBAR(who) → WHNP S	ADVP(currently) → RB	13	S(have) → NP VP .	VP(said) → VBD SBAR
SBAR(as) → IN S	VP(born) → VBN NP NP	14	S(could) → NP VP .	VP(has) → VBZ NP
WHNP(who) → WP	ADVP(initially) → RB	15	S(said) → NP VP .	**VP(is) → VBZ NP**
NP(i) → FW	**PP(with) → IN NP**	16	**S(has) → NP VP .**	NP(this) → DT
PP(as) → IN NP	WHPP(of) → IN WHNP	17	NP(people) → JJ NNS	VP(was) → VBD NP
NP(director) → NP PP	**SBAR(although) → IN S**	18	NP(money) → NN	**NP(people) → NNS**
PP(by) → IN NP	ADVP(primarily) → RB	19	NP(government) → DT NN	NP(lot) → DT NN
S(has) → VP	S(links) → NP VP .	20	S(do) → NP VP .	NP(season) → NN CD
PP(in) → IN NP	VP(links) → VBZ NP	21	NP(scientists) → NNS	**S(can) → NP VP .**
SBAR(while) → IN S	PP(following) → VBG NP	22	**VP(called) → VBN NP**	VP(is) → VBZ VP
PP(as) → JJ IN NP	ADVP(subsequently) → RB	23	S(had) → NP VP .	SBAR(because) → IN S
PRN(–) → : NP :	**SBAR(which) → WHNP S**	24	S(says) → NP VP .	VP(are) → VBP NP
S('s) → NP VP	**SBAR(while) → IN S**	25	S(would) → NP VP .	NP(player) → DT JJ NN NN
S(said) → " S , " NP VP .	S(plays) → ADVP VP	26	S(say) → NP VP .	NP(there) → EX
PP(at) → IN NP	PP(within) → IN NP	27	S(works) → NP VP .	NP(lot) → NP PP
PP(among) → IN NP	**PP(by) → IN NP**	28	S(may) → NP VP .	NP(websites) → JJ NNS
SBAR(although) → IN S	SBAR(of) → WHNP S	29	S(did) → NP VP .	PP(like) → IN NP
VP(said) → VBD NP	S(is) → S : S .	30	S(think) → NP VP .	S(started) → NP VP .

Table 9: Top 30 syntax patterns associated with the complex text (left) and simplified text (right). Bold patterns are the top patterns shared by Newsela and Wikipedia.

that are retained in Simple Wikipedia indicates the incompleteness of simplification in the Simple Wikipedia. The dramatic frequency decrease of words like "which" and "advocates" in Newsela shows the consistent quality from professional simplifications. Wikipedia has good coverage on certain words, such as "approximately", because of its large volume.

3.4 Log-odds-ratio analysis of syntax patterns

We can also reveal the syntax patterns that are most strongly associated with simple text versus complex text using the log-odds-ratio technique. Table 9 shows syntax patterns that represent "parent node (head word) → children node(s)" structures from a constituency parse tree. To extract theses patterns we parsed our corpus with the Stanford Parser (Klein and Manning, 2002) and applied its built-in head word identifier from Collins (2003). Both the Newsela and Wikipedia corpora exhibit syntactic differences that are intuitive and interesting. However, as with word frequency (Table 8), complex syntactic patterns are retained more often in Wikipedia's simplifications than in Newsela's.

In order to show interesting syntax patterns in the Wikipedia parallel data for Table 9, we first had to discard 3613 sentences in PWKP that contain both "is a commune" and "France". As the word-level analysis in Tables 6 and 7 hints, there is an exceeding number of sentences about communes in France in the PWKP corpus, such as the sentence pair below:

[NORM] *La Couture is a commune in the Pas-de-Calais department in the Nord-Pas-de-Calais region of France .*

[SIMP] *La Couture, Pas-de-Calais is a commune. It is found in the region Nord-Pas-de-Calais in the Pas-de-Calais department in the north of France.*

This is a template sentence from a stub geographic article and its deterministic simplification. The influence of this template sentence is more over-

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

whelming in the syntax-level analysis —- about 1/3 of the top 30 syntax patterns would be related to these sentence pairs if they were not discarded.

3.5 Document-level compression

There are few publicly accessible document-level parallel simplification corpora (Barzilay and Lapata, 2008). The Newsela corpus will enable more research on document-level simplification, such as anaphora choice (Siddharthan and Copestake, 2002), content selection (Woodsend and Lapata, 2011b), and discourse relation preservation (Siddharthan, 2003).

Simple Wikipedia is rarely used to study document-level simplification. Woodsend and Lapata (2011b) developed a model that simplifies Wikipedia articles while selecting their most important content. However, they could only use Simple Wikipedia in very limited ways. They noted that Simple Wikipedia is "less mature" with many articles that are just "stubs, comprising a single paragraph of just one or two sentences". We quantify their observation in Figure 2, plotting the document-level compression ratio of Simple vs. Normal Wikipedia articles. The compression ratio is the ratio of the number of characters between each simple-complex article pair. In the plot, we use all 60 thousand article pairs from the Simple-Normal Wikipedia collected by Kauchak (2013) in May 2011. The overall compression ratio is skewed towards almost 0. For comparison, we also plot the ratio between the simplest version (Simp-4) and the original version (Original) of the news articles in the Newsela corpus. The Newsela corpus has a much more reasonable compression ratio and is therefore likely to be more suitable for studying document-level simplification.

3.6 Analysis of discourse connectives

Although discourse is known to affect readability, the relation between discourse and text simplification is still under-studied with the use of statistical methods (Williams et al., 2003; Siddharthan, 2006; Siddharthan and Katsos, 2010). Text simplification often involves splitting one sentence into multiple sentences, which is likely to require discourse-level changes such as introducing explicit rhetorical relations. However, previous research that uses Simple-Normal Wikipedia largely focuses on sentence-level transformation, without taking large discourse structure into account.

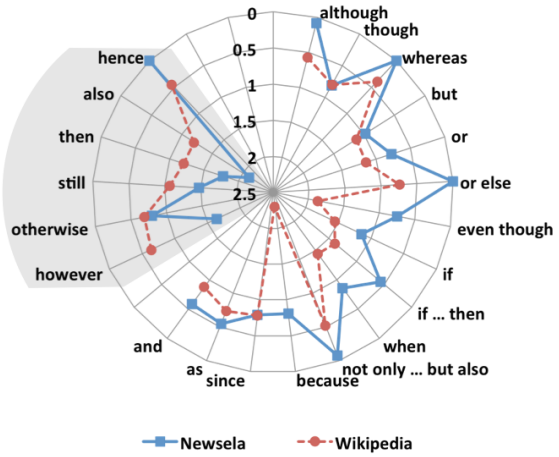

Figure 3: A radar chart that visualizes the odds ratio (radius axis) of discourse connectives in simple side vs. complex side. An odds ratio larger than 1 indicates the word is more likely to occur in the simplified text than in the complex text, and vice versa. Simple cue words (in the shaded region), except "hence", are more likely to be added during Newsela's simplification process than in Wikipedia's. Complex conjunction connectives (in the unshaded region) are more likely to be retained in Wikipedia's simplifications than in Newsela's.

To preserve the rhetorical structure, Siddharthan (2003, 2006) proposed to introduce cue words when simplifying various conjoined clauses. We perform an analysis on discourse connectives that are relevant to readability as suggested by Siddharthan (2003). Figure 3 presents the odds ratios of simple cue words and complex conjunction connectives. The odds radios are computed for Newsela between the Original and Simp-4 versions, and for Wikipedia between Normal and Simple documents collected by Kauchak (2013). It suggests that Newsela exhibits a more complete degree of simplification than Wikipedia, and that it may be able to enable more computational studies of the role of discourse in text simplification in the future.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

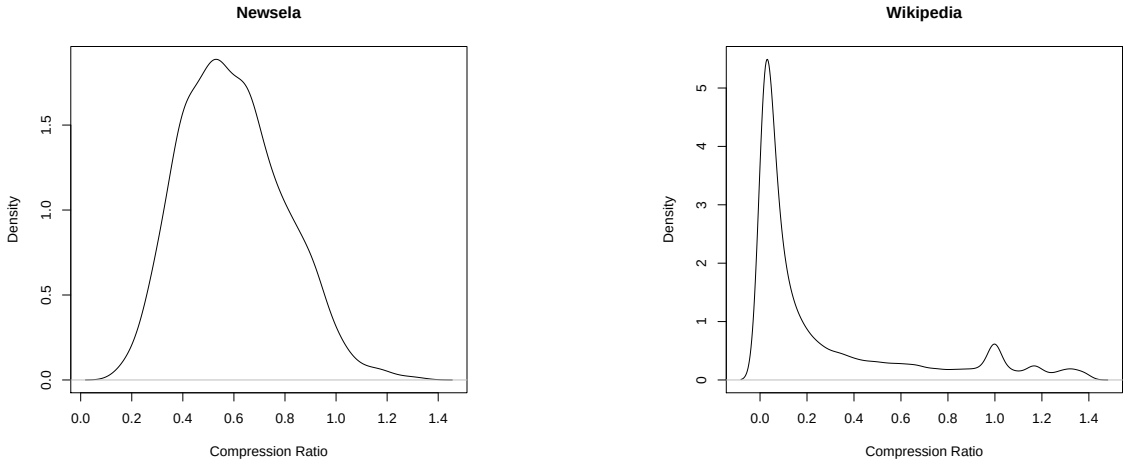

Figure 2: Distribution of document-level compression ratio, displayed as a histogram smoothed by kernel density estimation. The Newsela corpus is more normally distributed, suggesting more consistent quality.

3.7 Newsela's quality is better than Wikipedia

Overall, we have shown that the professional simplification of Newsela is more rigorous and more consistent than Simple English Wikipedia. The language and content also differ between the encyclopedia and news domains. They are not exchangeable in developing nor in evaluating simplification systems. In the next section, we will review the evaluation methodology used in recent research, discuss its shortcomings and propose alternative evaluations.

4 Evaluation of simplification systems

With the popularity of parallel Wikipedia data in simplification research, most state-of-the-art systems evaluate on simplifying sentences from Wikipedia. All simplification systems published in the ACL, NAACL, EACL, COLING and EMNLP main conferences since Zhu's 2010 work compared solely on the same test set that consists of only 100 sentences from Wikipedia, except one paper that additionally experimented with 5 short news summaries. The most widely practiced evaluation methodology is to have human judges rate on grammaticality (or fluency), simplicity, and adequacy (or meaning preservation) on a 5-point Likert scale.

Such evaluation is insufficient to measure 1) the practical value of a system to a specific target reader population and 2) the performance of individual simplification components: sentence splitting, deletion and paraphrasing. Although the inadequacy of text simplification evaluations has been discussed before (Siddharthan, 2014), we focus on these two common deficiencies and suggest two future directions.

4.1 Targeting specific audiences

Simplification has many subtleties, since what constitutes simplification for one type of user may not be appropriate for another. Many researchers have studied simplification in the context of different audiences. However, most recent automatic simplification systems are developed and evaluated with little consideration of target reader population. There is one attempt by Angrosh et al. (2014) who evaluate their system by asking non-native speakers comprehension questions. They conducted an English vocabulary size test to categorize the users into different levels of language skills.

The Newsela corpus allows us to target children at different grade levels. From the application point of view, making knowledge accessible to all children is an important yet challenging part of education (Scarton et al., 2010; Moraes et al., 2014). From the technical point of view, reading grade level is a clearly defined objective for both simplification systems and human annotators. Once there is a well-defined objective, with constraints such as vocabulary size and sentence length, it is easier to fairly compare different systems. Newsela provides human simplification

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

at different grade levels and reading comprehension quizzes alongside each article.

In addition, readability is widely studied and can be automatically estimated (Kincaid et al., 1975; Pitler and Nenkova, 2008; Petersen and Ostendorf, 2009). Although existing readability metrics assume text is well-formed, they can potentially be used in combination with text quality metrics (Post, 2011; Louis and Nenkova, 2013) to evaluate simplifications. They can also be used to aid humans in the creation of reference simplifications.

4.2 Evaluating sub-tasks separately

It is widely accepted that sentence simplification involves three different elements: *splitting*, *deletion* and *paraphrasing* (Feng, 2008; Narayan and Gardent, 2014). Splitting breaks a long sentence into a few short sentences to achieve better readability. Deletion reduces the complexity by removing unimportant parts of a sentence. Paraphrasing rewrites text into a simpler version via reordering, substitution and occasionally expansion.

Most state-of-the-art systems consist of all or a subset of these three components. However, the popular human evaluation criteria (grammaticality, simplicity and adequacy) do not explain which components in a system are good or bad. More importantly, deletion may be unfairly penalized since shorter output tends to result in lower adequacy judgements (Napoles et al., 2011).

We therefore advocate for a more informative evaluation that separates out each sub-task. We believe this will lead to more easily quantifiable metrics and possibly the development of automatic metrics. For example, early work shows potential use of precision and recall to evaluate splitting (Siddharthan, 2006; Gasperin et al., 2009) and deletion (Riezler et al., 2003; Filippova and Strube, 2008). Several studies also have investigated various metrics for evaluating sentence paraphrasing (Callison-Burch et al., 2008; Chen and Dolan, 2011; Ganitkevitch et al., 2011; Xu et al., 2012, 2013; Weese et al., 2014).

5 Summary and recommendations

In this paper, we presented the first systematic analysis of the quality of Simple Wikipedia as a simplification data resource. We conducted a qualitative manual examination and several statistical analyses (including vocabulary change matrices, compression ratio histograms, log-odds-ratio calculations, etc.). We introduced a new, high-quality corpus of professionally simplified news articles, Newsela, as an alternative resource, that allowed us to demonstrate Simple Wikipedia's inadequacies in comparison. We further discussed problems with current simplification evaluation methodology and proposed potential improvements.

Our goal for this opinion paper is to stimulate progress in text simplification research. Simple English Wikipedia played a vital role in inspiring simplification approaches based on statistical machine translation. However, it has so many drawbacks that we recommend the community to drop it as the standard benchmark set for simplification. Other resources like the Newsela corpus are superior, since they provide a more consistent level of quality, target a particular audience, and approach the size of parallel Simple-Normal English Wikipedia. We believe that simplification is an important area of research that has the potential for broader impact beyond NLP research. But we must first adopt appropriate data sets and research methodologies.

Researchers can request the Newsela data following the instructions at: `https://newsela.com/data/`

Acknowledgments

The authors would like to thank Dan Cogan-Drew, Jennifer Coogan, and Kieran Sobel from Newsela for creating their data and generously sharing it with us. We also thank action editor Rada Mihalcea and three anonymous reviewers for their thoughtful comments, and Ani Nenkova, Alan Ritter and Maxine Eskenazi for valuable discussions.

This material is based on research sponsored by the NSF under grant IIS-1430651. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of the NSF or the U.S. Government.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

References

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599.

Angrosh, M., Nomoto, T., and Siddharthan, A. (2014). Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Bach, N., Gao, Q., Vogel, S., and Waibel, A. (2011). Tris: A statistical sentence simplifier with log-linear models and margin-based discriminative training. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Callison-Burch, C., Cohn, T., and Lapata, M. (2008). ParaMetric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*.

Canning, Y., Tait, J., Archibald, J., and Crawley, R. (2000). Cohesive generation of syntactically simplified newspaper text. In *Proceedings of the Third International Workshop on Text, Speech and Dialogue (TSD)*.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 14th Conference of the 9th European Conference for Computational Linguistics (EACL)*.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational linguistics (COLING)*.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chen, H.-B., Huang, H.-H., Chen, H.-H., and Tan, C.-T. (2012). A simplification-translation-restoration framework for cross-domain smt applications. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637.

Coster, W. and Kauchak, D. (2011). Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

De Belder, J. and Moens, M.-F. (2010). Text simplification for children. In *Prroceedings of the SIGIR Workshop on Accessible Search Systems*.

Eisenstein, J. (2013). What to do about bad language on the Internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Elhadad, N. and Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*.

Feng, L. (2008). Text simplification: A survey. Technical report, The City University of New York.

Filippova, K. and Strube, M. (2008). Dependency tree based sentence compression. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gasperin, C., Maziero, E., Specia, L., Pardo, T., and Aluisio, S. M. (2009). Natural language processing for social inclusion: A text simplification architecture for different literacy levels. In *Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*.

Gerber, L. and Hovy, E. (1998). Improving translation quality by manipulating sentence length. In

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Machine Translation and the Information Soup, pages 448–460. Springer.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: A project note. In *Proceedings of the 2nd International Workshop on Paraphrasing*.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Jurafsky, D., Chahuneau, V., Routledge, B. R., and Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4).

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL)*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Klebanov, B. B., Knight, K., and Marcu, D. (2004). Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 735–747. Springer.

Klein, D. and Manning, C. D. (2002). Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*.

Louis, A. and Nenkova, A. (2013). What makes writing great? First experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics (TACL)*, 1:341–352.

Miwa, M., Saetre, R., Miyao, Y., and Tsujii, J. (2010). Entity-focused sentence simplification for relation extraction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Moraes, P., McCoy, K., and Carberry, S. (2014). Adapting graph summaries to the users' reading levels. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*.

Napoles, C., Callison-Burch, C., and Van Durme, B. (2011). Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*.

Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Petersen, S. and Ostendorf, M. (2007). Text simplification for language learners: A corpus analysis. In *Proceedings of the Workshop on Speech and Language Technology for Education*.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Porter, A., McMaken, J., Hwang, J., and Yang, R. (2011). Common Core Standards the new US intended curriculum. *Educational Researcher*, 40(3):103–116.

Post, M. (2011). Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Riezler, S., King, T. H., Crouch, R., and Zaenen, A. (2003). Statistical sentence condensation us-

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

ing ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*.

Scarton, C., De Oliveira, M., Candido Jr, A., Gasperin, C., and Aluísio, S. M. (2010). Simplifica: A tool for authoring simplified texts in brazilian portuguese guided by readability assessments. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Siddharthan, A. (2003). Preserving discourse structure when simplifying text. In *Proceedings of European Workshop on Natural Language Generation (ENLG)*.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Siddharthan, A. (2014). A survey of research on text simplification. *Special issue of International Journal of Applied Linguistics*, 165(2).

Siddharthan, A. and Angrosh, M. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*.

Siddharthan, A. and Copestake, A. (2002). Generating anaphora for simplifying text. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*.

Siddharthan, A. and Katsos, N. (2010). Reformulating discourse connectives for non-expert readers. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Siddharthan, A., Nenkova, A., and McKeown, K. (2004). Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Vickrey, D. and Koller, D. (2008). Sentence simplification for semantic role labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Aluísio, S. M. (2009). Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication*.

Weese, J., Ganitkevitch, J., and Callison-Burch, C. (2014). PARADIGM: Paraphrase diagnostics through grammar matching. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Williams, S., Reiter, E., and Osman, L. (2003). Experiments with discourse-level choices and readability. In *Proceedings of the European Natural Language Generation Workshop (ENLG)*.

Woodsend, K. and Lapata, M. (2011a). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Woodsend, K. and Lapata, M. (2011b). WikiSimple: Automatic simplification of Wikipedia articles. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*.

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 (ACL)*.

Xu, W. and Grishman, R. (2009). A parse-and-trim approach with information significance for chinese sentence compression. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Xu, W., Ritter, A., Dolan, B., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

Xu, W., Ritter, A., and Grishman, R. (2013). Gathering and generating paraphrases from twitter with application to normalization. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora (BUCC)*.

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 (COLING)*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00139 by guest on 12 April 2021