Text Simplification from Professionally Produced Corpora

Carolina Scarton, Gustavo Henrique Paetzold, Lucia Specia
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{g.h.paetzold, c.scarton, l.specia}@sheffield.ac.uk

Abstract

The lack of large and reliable datasets has been hindering progress in Text Simplification (TS). We investigate the application of the recently created Newsela corpus, the largest collection of professionally written simplifications available, in TS tasks. Using new alignment algorithms, we extract $550,644$ complex-simple sentence pairs from the corpus. This data is explored in different ways: (i) we show that traditional readability metrics capture surprisingly well the different complexity levels in this corpus, (ii) we build machine learning models to classify sentences into complex vs. simple and to predict complexity levels that outperform their respective baselines, (iii) we introduce a lexical simplifier that uses the corpus to generate candidate simplifications and outperforms the state of the art approaches, and (iv) we show that the corpus can be used to learn sentence simplification patterns in more effective ways than corpora used in previous work.

Keywords: text Simplification, simplification corpora, Newsela

1. Introduction

Text Simplification (TS) consists in making texts more easily comprehensible. It can take many forms: Lexical Simplification (LS), in which complex words are replaced by simpler alternatives (Devlin, 1999), Syntactic Simplification (SS), which consists in changing the syntactic structure of a sentence (Siddharthan, 2006), and Semantic Simplification, in which portions of the text are paraphrased (Kandula et al., 2010).

Current empirical approaches rely mostly on the Wikipedia-Simple Wikipedia parallel corpus (Coster and Kauchak, 2011). This resource has been used by Machine Translation (MT) approaches (Zhu et al., 2010), tree transductors (Paetzold and Specia, 2013; Feblowitz and Kauchak, 2013), integer programming techniques (Woodsend and Lapata, 2011), and discriminative linear models (Bach et al., 2011). In LS, Yatskar et al. (2010) extract candidate simplifications from Wikipedia and Simple Wikipedia edit histories, and Horn et al. (2014) extract word correspondences from word alignments between the complex-simple segments in the corpus.

Even though the Simple Wikipedia corpus has been a valuable resource for modern TS, as discussed in Yasseri et al. (2012), Amancio and Specia (2014) and Xu et al. (2015), it is very small ($167,689$ parallel sentence pairs (Coster and Kauchak, 2011)) in comparison to bilingual corpora used with similar modelling techniques in MT and, more critically, covers a limited range of simplification operations, which are applied in *ad hoc* ways by volunteer editors. Xu et al. (2015) introduce a new resource that allegedly addresses these limitations: the **Newsela corpus** (Newsela, 2016). Unlike Simple Wikipedia, the Newsela corpus was created by professional editors and targets a specific audience (students), which should make it a more reliable resource for TS. However, the Newsela corpus has only recently started to be exploited for this task and not enough work has been done to understand its potential.

In this paper, we investigate whether (Xu et al., 2015)'s claims hold in practice. We produce sentence alignments for the Newsela corpus (Section 2.) and conduct experiments to evaluate its effectiveness in TS tasks, namely: readability analysis (Section 4.), complex vs. simple classification, complexity level prediction, lexical simplification and MT-based sentence simplification (Section 5.).

2. Aligning the Newsela Corpus

The Newsela corpus (version 2016-01-29.1) is composed of $10,787$ news articles in English, which includes $1,911$ articles in their original form as well as in 4 or 5 versions rewritten by humans to suit different reading reading levels. Each document is characterised by a unique identifier, a version identifier (from 0 – most complex to 5 – simplest), and a reading level from 2 to 12, where 2 represents the lowest and 12 the highest level. Version identifiers capture the relationship between the reading levels of a pair of documents: e.g. version 1 has a lower reading level than version 0, version 2 has a lower reading level than version 1.

Articles are only aligned at document level and there is no guarantee that different versions of an article will have the same number of sentences, nor that they will be aligned in 1-to-1 fashion. The absence of paragraph and sentence alignments limits the use of the data.

To produce such alignments, we use the algorithms in (Paetzold and Specia, 2016d), which employ a vicinity-driven search approach. These algorithms address the limitations of previous strategies (Barzilay and Elhadad, 2003; Coster and Kauchak, 2011; Smith et al., 2010; Xu et al., 2015; Bott and Saggion, 2011) by disregarding the need for supervised or semi-supervised training, allowing long-distance alignment skips, capturing 1-N and N-1 alignments, and exploiting the fact that the order in which information is presented is constant between pairs of aligned Newsela articles. Because the vicinity-driven approach of Paetzold and Specia (2016d) exploits a series of assumptions that can be made about the Newsela corpus, it is more efficient than more sophisticated approaches that perform exhaustive search over all possible paragraph/sentence alignments (Štajner et al., 2017), while still offering comparable alignment accuracy.

The result of the alignment is a corpus with 19,198 pairs of articles aligned at both paragraph (300,475 pairs) and sentence (550,644 pairs) levels. This is over three times larger than the Wikipedia–Simple Wikipedia corpus (Coster and Kauchak, 2011), making it the largest corpus of its kind.

Columns 2 to 4 in Table 1 illustrate the number of paragraph and sentence alignments for all version pairs in the corpus. We categorise the sentence alignments according to four types of simplification:

- **None:** Complex and simple sentences are identical (146,251 pairs).
- **Compression:** Multiple complex sentences are aligned to fewer simple sentences (24,661 pairs).
- **Splitting:** Multiple simple sentences are aligned to fewer complex sentences: (121,582 pairs).
- **Rewriting:** Same number of complex and simple sentences, but with different content: (258,150 pairs).

Pair	# Doc.	# Parag.	# Sent.	% None	Avg. TER
0-1	1,910	39,414	69,443	42.1	**0.193**
0-2	1,910	35,720	60,725	26.5	0.316
0-3	1,910	27,752	44,168	16.7	0.449
0-4	1,882	19,369	28,499	12.3	0.537
0-5	42	261	346	5.5	0.647
1-2	1,910	38,497	75,953	37.8	**0.222**
1-3	1,910	30,824	55,572	19.1	0.400
1-4	1,882	22,163	36,089	12.3	0.511
1-5	42	300	417	5.3	0.651
2-3	1,910	33,033	69,416	29.4	**0.308**
2-4	1,882	24,363	45,392	15.8	0.455
2-5	42	329	523	6.5	0.605
3-4	1,882	27,635	62,413	29.6	**0.325**
3-5	42	386	706	9.5	0.554
4-5	42	429	982	20.5	**0.423**
Total	19,198	300,475	550,644	26.6	0.440

Table 1: Documents, aligned paragraphs and sentences at all levels (columns 2-4), % of "none" alignments, and average TER (columns 5-6)

On average, sentence lengths remain close for adjacent levels (e.g. 25.0 & 24.4 for levels 0-1), but sentences become shorter for higher levels (12.5 & 11.1 at levels 4-5). This shows that editors significantly compress text while simplifying.

3. Related Work

Xu et al. (2015) are the first to present an analysis of the Newsela corpus. They compare the Newsela and Wikipedia-Simple Wikipedia data using several metrics, showing that the Newsela corpus appears to be more useful. However, they do not use it in any tasks (e.g. lexical simplification) like we propose in this paper.

Besides proposing alignment algorithms for the Newsela corpus, Štajner et al. (2017) also build MT-based models with the aligned data. As test set, instead of using part of Newsela data, the Wikipedia-Simple Wikipedia dataset proposed by Xu et al. (2016) is used. Two out of three systems trained with Newsela aligned data perform better in terms of simplicity than state-of-the-art systems for the same corpus.

Zhang and Lapata (2017) train an attention-based encoder-decoder model (Bahdanau et al., 2014) and use reinforcement learning with a reward policy combining SARI (to measure simplicity) (Xu et al., 2016), BLEU (to measure grammaticality) (Papineni et al., 2002) and cosine similarity (to measure meaning preservation). This approach shows improvements over a model trained using a phrase-based MT approach in terms of BLEU and SARI.

Alva-Manchego et al. (2017) propose a TS model that uses predicted simplification operations. Simplification operations automatically annotated in source sentences are predicted as a first step, using sequence labelling techniques. As operations, they model only *replace* and *delete*. Their TS model produces better results according to human judgements for simplicity than general-purpose MT-based models.

In general, the aforementioned contributions explore MT-based techniques and train systems using the Newsela data in similar ways as it was done previously for Wikipedia-Simple Wikipedia data. However, none of them provide an analysis of the Newsela data in terms of readability of the aligned data, the use of the data for complex vs. simple classification or complexity level prediction, or the impact of the data in state-of-the-art LS approaches.

4. Corpus Analysis

We analyse the sentence-aligned Newsela corpus to (i) understand the differences between its various levels of simplification, and (ii) investigate how existing readability and psycholinguistic metrics fair in distinguishing these levels.

Edit Rate This analysis focuses on the differences in edits between the various simplified versions. We use TER[1] as a metric of edit distance, as it is widely used for this purpose in MT evaluation.

Columns 5 and 6 of Table 1 show the percentage of alignments with TER $= 0$ ("% None") and averaged TER. As expected, the non-adjacent versions have higher TER values, e.g. the distance between levels '0' and '1' is much smaller than the distance between levels '0' and '5'. The percentage of sentences with no edits decreases as we move from adjacent to non-adjacent levels. Interestingly, between the adjacent levels, the closer to the original level, the lower the TER, e.g. there are fewer edits between '0' and '1' than between '1' and '2'.

Readability Metrics Here we evaluate standard readability metrics that aggregate shallow text information (such as number of syllables and words): Flesch Reading Ease, Flesch-Kincaid Grade Level, SMOG Index, Gunning Fog Index, Automated Readability Index, Coleman-Liau Index, Linsear Write Formula and Dale-Chall Readability Score from the TEXTSTAT toolkit[2]. Figure 1 shows the Flesch Reading Ease box plot for pairs of original and simplified sentences with level '0' as original version. Flesch varies from 0 (most complex) to 100 (simplest). The Flesch index for the simplified versions is higher than for the original version in all cases, which is an expected behaviour. Therefore, although the simplified versions of the Newsela

[1] `http://www.cs.umd.edu/~snover/tercom`

[2] `https://pypi.python.org/pypi/textstat`

corpus can be composed of more and/or longer sentences than the original, the information encoded in them is still simpler.

A similar trend is observed for all other readability metrics and between all levels. For completion, we also show in Figure 2 the box plot for the Flesch-Kincaid Grade Level metric for 0-to-n original/simplified pairs. As expected, simplified versions have lower Flesch-Kincaid scores.

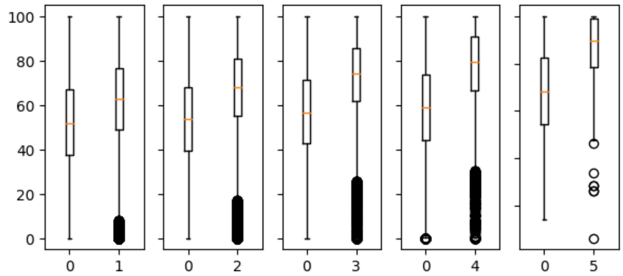

Figure 1: Flesch Reading Ease for simplified versions from level ‘0’

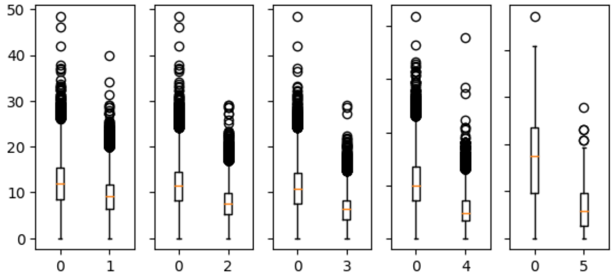

Figure 2: Flesch-Kincaid Grade Level for simplified versions from level ‘0’

Psycholinguistic Metrics We also analysed the data using 12 psycholinguistic features, motivated by those in the Coh-Metrix tool (Graesser et al., 2004):[3]

- Number of words / tokens / letters/ syllables
- Type/token ratio
- Ratio betw. numbers of letters and words
- Ratio betw. numbers of syllables and words
- Number of content words
- Mean age of acquisition / familiarity / imageability / concreteness score of content words

For sentences without content words, the correspondent features were assigned zero. We extracted age of acquisition, familiarity, imageability and concreteness features from the bootstrapped MRC database (Paetzold and Specia, 2016b), which is an extended version of the original MRC database (Coltheart, 1981).

Original sentences have a higher number of words, tokens, letters and syllables, as expected. Type/token ratio and number of content words are not considerably different between original and simplified versions. Figures 3 and 4 show the box plots for the ratio between number of letter and number words and the ration between number of syllables and number words, respectively, when ‘0’ is the original level. Simplified versions have a higher ratio of letters per words and a lower ratio of syllables per words, when compared to original versions. It appears that even though simplified sentences have slightly longer words, such words have fewer syllables, which is often a sign of simplicity.

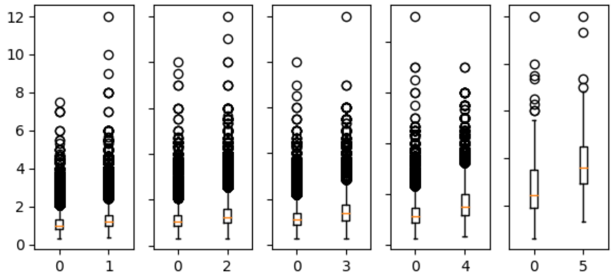

Figure 3: Ratio between number of letters and number of words for simplified versions from level ‘0’

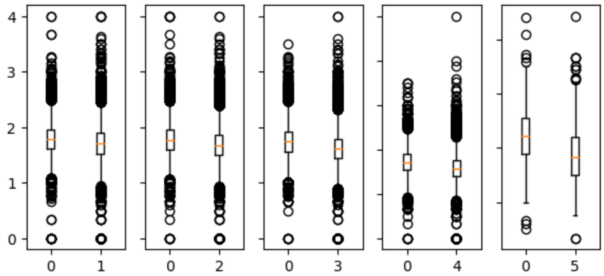

Figure 4: Ratio between number of syllables and number of words for simplified versions from level ‘0’

Figures 5 and 6 show the box plots for age of acquisition and imageability metrics, when ‘0’ is the original. Age of acquisition aims to define the age at which a given word is learned, whilst imageability refers to the mental capability of retrieving an image, given a word. As expected, simplified sentences show lower values for age of acquisition and higher values for imageability than their original counterparts. Familiarity (the frequency to which a word is seen, heard or spoken daily) and concreteness did not show differences between simplified and original sentences.

5. Using the Corpus in TS Tasks

5.1. Complex vs. Simple Classification

Here we present sentence-level binary classifiers created for all possible combinations of levels of simplification. Sentences from the more complex version were assigned the label “complex”, while their simpler counterpart (at any level), the label “simple”. For sentences pairs whose TER is 0 (no simplification made), both original and simplified sentences were considered “simple”.

We trained Stochastic Gradient Descent (SGD) classifiers (with hinge loss function) using the scikit-learn toolkit (Pe-

[3] http://cohmetrix.com

	Majority			Readability			Psycholinguistic			All		
	P	R	*F*	P	R	*F*	P	R	*F*	P	R	*F*
0-1	0.634	1.0	0.491	0.684	0.929	0.641	0.651	0.889	0.576	0.696	0.936	**0.661**
0-2	0.576	1.0	0.421	0.715	0.857	0.711	0.660	0.826	0.633	0.724	0.860	**0.724**
0-3	0.545	1.0	0.385	0.740	0.875	0.759	0.687	0.782	0.684	0.753	0.872	**0.770**
0-4	0.533	1.0	0.370	0.756	0.917	0.794	0.708	0.789	0.717	0.776	0.909	**0.808**
0-5	0.514	1.0	0.349	0.758	0.951	0.809	0.718	0.837	0.740	0.767	0.953	**0.823**
1-2	0.616	1.0	0.470	0.681	0.881	0.652	0.643	0.881	0.555	0.696	0.868	**0.664**
1-3	0.553	1.0	0.393	0.715	0.853	0.725	0.674	0.752	0.664	0.727	0.845	**0.734**
1-4	0.533	1.0	0.370	0.744	0.891	0.778	0.713	0.773	0.704	0.765	0.890	**0.791**
1-5	0.514	1.0	0.349	0.798	0.954	0.849	0.742	0.781	0.750	0.801	0.957	**0.853**
2-3	0.586	1.0	0.433	0.662	0.861	0.638	0.633	0.896	0.584	0.674	0.848	**0.653**
2-4	0.543	1.0	0.382	0.704	0.851	0.717	0.677	0.743	0.664	0.722	0.847	**0.732**
2-5	0.517	1.0	0.352	0.761	0.929	0.800	0.726	0.776	0.730	0.772	0.922	**0.810**
3-4	0.587	1.0	0.434	0.649	0.883	0.611	0.622	0.828	0.551	0.655	0.869	**0.628**
3-5	0.525	1.0	0.361	0.709	0.875	0.727	0.685	0.729	0.673	0.711	0.888	**0.745**
4-5	0.557	1.0	0.399	0.664	0.837	0.656	0.644	0.753	0.626	0.644	0.789	**0.622**

Table 2: Precision, recall and F-measure of classifiers for different simplification levels

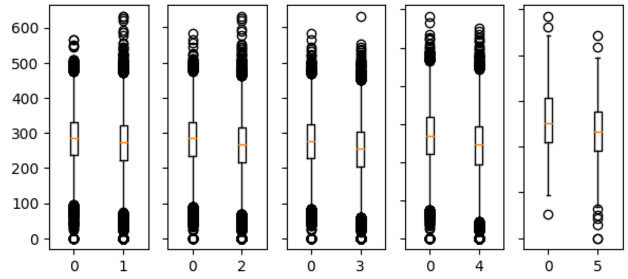

Figure 5: Age of acquisition for simplified versions from level ‘0’

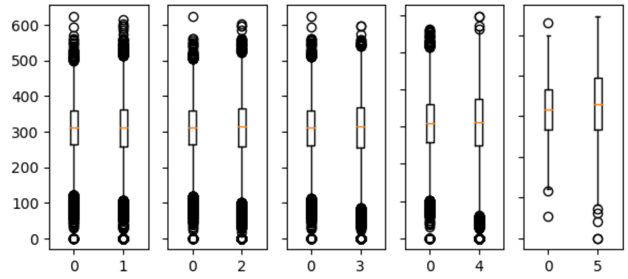

Figure 6: Imageability for simplified versions from level ‘0’

dregosa et al., 2011) with hyperparameters optmised using grid search. As features we used the nine readability metrics from the TEXTSTAT toolkit and the 12 psycholinguistic features, mentioned in Section 4..

Three models were built: one with the readability metrics only, one with the psycholinguistic metrics only and another with both. The classifiers were evaluated by using 10-fold cross-validation. As a baseline, we used a majority class classifier (Majority).

Table 2 shows the results for each complex-simple level. As expected, the classifiers built for non-adjacent levels achieve better performance than those for adjacent levels (in terms of F-measure). The opposite behaviour is observed for the majority class models. This is expected, however, since the number original/simplified pairs whose TER is 0 is much larger in adjacent levels than in non-adjacent levels, leading to more biased (and hence easier to predict) instances. It can also be noticed that the precision and F-measure of the classifiers follow the degree of difference between the complex and simple levels, as shown by TER (Table 1). Recall, on the other hand, is higher for adjacent levels. This is most likely caused also by the large number of sentences considered simple because of no changes in TER (% None in Table 1). All classifiers outperform the majority class baseline and the best classifiers use the combination of both types of metrics as feature.

5.2. Complexity Level Prediction

Here we directly predict the level of complexity of a sentence. These are defined as 2-12 reading proficiency levels, as explained in Section 2.: the higher the level, more complex the text. We use the same feature sets as in Section 5.1., but mix all sentences to build a single model and use the Ridge Regression algorithm.

We evaluate the model in terms of Mean Absolute Error (MAE). As baseline, we considered the MAE obtained from applying the mean complexity level of the training set as the “prediction” for all instances in the test set. MAE values of 1.793, 1.962 and 1.715 were obtained for models built with readability, psycholinguistic and all features, respectively. The mean baseline MAE was of 2.247, and therefore all models outperformed the baseline, with the best model using all features.

5.3. Lexical Simplification

Generator	Pot.	Prec.	Rec.	F1
Horn	0.569	0.235	0.131	0.168
Devlin	0.647	0.133	0.153	0.143
Biran	0.610	0.130	0.144	0.136
Glavas	0.724	0.142	0.191	0.163
Paetzold	**0.856**	0.180	**0.252**	**0.210**
Newsela	0.602	**0.304**	0.128	0.180

Table 3: Substitution Generation results

	Famil.	Glavas	Paetzold
Horn	0.334	0.332	0.352
Devlin	0.291	0.265	0.341
Biran	0.230	0.267	0.238
Glavas	0.186	0.251	0.231
Paetzold	0.350	0.325	0.378
Newsela	**0.372**	**0.344**	**0.400**

Table 4: Accuracy in the full pipeline evaluation

Here we assess the potential of our corpus in LS. LS is commonly addressed as a pipeline of steps: candidates for a target complex word are produced via a **Substitution Generation** (SG) method, filtered with respect to the context of the complex word via a **Substitution Selection** (SS) method, and finally ordered for simplicity by a **Substitution Ranking** (SR) method.

We use our aligned corpus for SG following the state of the art approach in (Horn et al., 2014). First, we produce word alignments using Meteor (Denkowski and Lavie, 2011) and extract complex-to-simple word correspondences. Then we filter word pairs with different POS tags, where the complex word is a stop word, or either word is a proper noun. Finally, we generate all possible inflections for nouns and verbs (Burns, 2013).

We compare this approach to six other generators from a recent benchmark (Paetzold and Specia, 2016a): the Horn generator (Horn et al., 2014), which employs the approach described above over Wikipedia-Simple Wikipedia data, the Devlin (Devlin and Tait, 1998), Biran (Biran et al., 2011), Glavas (Glavaš and Štajner, 2015) and Paetzold (Paetzold and Specia, 2016c) generators, which exploit WordNet, comparable complex-to-simple documents, typical word embeddings and context-aware word embeddings, respectively. All generators were implemented with the LEXenstein framework (Paetzold and Specia, 2015).

We use the BenchLS dataset as our gold-standard dataset (Paetzold and Specia, 2016a). It is the largest dataset of its kind, with 929 instances, each composed by a sentence, a target complex word, and a set of gold substitutions given by humans. To compare the generators, we use standard metrics: **Potential** – the proportion of instances in which at least one of the candidates generated is in the gold-standard, **Precision** – the proportion of generated substitutions that are in the gold-standard, **Recall** – the proportion of gold-standard substitutions that are among the generated substitutions, and **F1**. Table 3 reveals that our approach achieves the highest Precision overall, as well as higher Potential and F1 scores.

We also evaluated our generator in practice through a full pipeline evaluation, where the output is the best lexical simplification for each complex word. To do so, we paired all aforementioned generators with three state of the art SR strategies:

- **Familiarity** (Paetzold and Specia, 2016b): Ranks candidates according to their word familiarity scores, as extracted from the bootstrapped MRC database.
- **Glavas** (Glavaš and Štajner, 2015): Ranks candidates according to various features, then obtains a final ranking for a candidate by averaging the ranks of said features.
- **Paetzold** (Paetzold and Specia, 2015): Learns a ranking model from a binary classification setup.

All rankers were implemented with the LEXenstein framework with features and settings as in (Paetzold and Specia, 2016a). The gold-standard test set used is also BenchLS, and the metric is **Accuracy**: the ratio with which the highest ranking candidate is not the target word itself and is among the gold-standard candidates. Table 4 shows that our generator outperformed all others with any ranking method, highlighting the potential of the Newsela corpus for LS.

5.4. Sentence Simplification

Simplification can be addressed as a "translation approach" (Shardlow, 2014). This approach requires a large enough sentence-aligned complex-simple corpus and a method to learn simplification rules, such as off-the-shelf Statistical Machine Translation (SMT) toolkits.

We experiment with the aligned Newsela corpus following an SMT-like pipeline, using only the adjacent levels of simplification (0-1, 1-2, 2-3, 3-4 and 4-5) (278.184 sentences). For comparison, we build a model using the Wikipedia-Simple Wikipedia sentence-aligned corpus ($167,689$ sentences). Both datasets were divided in approximately 70% for training, 10% for development and 20% for test. Additionally, we also built a model using a subset of the Newsela dataset containing the same number of sentence pairs as the Wikipedia-Simple Wikipedia dataset, i.e. the training and development sets were reduced via random sampling to the size of the Wikipedia-Simple Wikipedia dataset.

We train standard MOSES toolkit (Koehn et al., 2007) with default configurations. Table 5 shows the evaluation of the simplification systems built in terms of BLEU, SARI and Flesch Ease Index. In this table "full" refers to the model trained with the entire Newsela dataset and "wiki size" refers to the model trained with the portion of the Newsela dataset with the same size as the Wikipedia-Simple Wikipedia dataset.

BLEU scores are higher for the Newsela trained system in both "full" and "wiki size" settings, which can indicate that the models trained with Newsela data are producing outputs more grammatically correct than the model trained with Wikipedia-Simple Wikipedia data. SARI, however, shows that models built with Wikipedia-Simple Wikipedia data seems to be producing slightly simpler outputs. The Flesch index for simplified sentences (FLESCH-S) is lower than that of the reference sentences (FLESCH-R) for both corpora. This seems to reflect the fact that automatic simplifications are closer to the original (FLESCH-O) than to the reference.

We also experiment with variants of the test set where only sentences with at least one edit (TER > 0) or with no edits at all (TER $= 0$) are used. Sentences with TER $= 0$ should not be modified as they are already simple, and thus the MT output should be exactly the same as the reference (and the original). This is often a problem in SMT-based simplification approaches, which tend to over simplify and introduce noise. The Newsela trained models are still the best in terms of BLEU, while show slightly smaller SARI. However, as the results for all TER $= 0$ suggest, SARI is

	BLEU	SARI	FLESCH-S	FLESCH-O	FLESCH-R
Wikipedia	0.569	0.302	66.93	66.21	74.32
Wikipedia: TER = 0	0.918	0.330	71.12	69.85	69.85
Wikipedia: TER > 0	0.454	0.292	67.29	65.34	79.98
Newsela (full)	0.692	0.270	75.04	74.89	80.62
Newsela (full): TER = 0	0.992	0.330	87.49	87.33	87.33
Newsela (full): TER > 0	0.575	0.238	70.34	70.19	78.91
Newsela (wiki size)	0.691	0.272	75.10	74.89	80.62
Newsela (wiki size): TER = 0	0.991	0.330	87.52	87.33	87.33
Newsela (wiki size): TER > 0	0.574	0.240	70.41	70.19	78.91

Table 5: Results for SMT-based simplifiers

not a reliable metric when original, reference and simplified sentences are the same. For all cases where TER = 0, the SARI value was 0.330, which can be seem as a low value if the systems are producing an output equal to the reference. Since this metric was designed for cases where sentences should also be simplified (as explained in Xu et al. (2016)), the use of SARI for cases where the original sentences are already simple is not reliable.

6. Conclusions

Upon studying the sentence-aligned Newsela corpus we found that: (i) it follows an expected TER distribution, with the lowest TER being between adjacent levels; (ii) the simplified sentences score as more readable than their original counterparts according to traditional readability metrics, and (iii) the corpus proved a more reliable source of complex-simple correspondences for LS and MT-based simplification than the Wikipedia-Simple Wikipedia corpus. We achieve some the highest performance to date when generating candidate substitutions for complex words as well as when applying these into a full LS pipeline. Improvements for MT-based simplification using the Newsela corpus are also observed but more in depth (manual) evaluation is needed for these experiments.

In the future, we hope that the aligned corpus will lead to better data-driven approaches to TS. We cannot release the aligned Newsela corpus, but it can be recreated using MASSAlign[4] (Paetzold et al., 2017), which provides the alignment algorithm used.

7. Acknowledgements

This work was supported by the EC project SIMPATICO (H2020-EURO-6-2015, grant number 692819).

8. Bibliographical References

Alva-Manchego, F., Bingel, J., Paetzold, G. H., Scarton, C., and Specia, L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of IJCNP*, pages 295–305.

Amancio, M. A. and Specia, L. (2014). An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd PITR*, pages 123–130.

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 *IJCNLP*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 EMNLP*, pages 25–32.

Biran, O., Brody, S., and Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th ACL*, pages 496–501.

Bott, S. and Saggion, H. (2011). An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the 2011 MTTG*, pages 20–26.

Burns, P. R. (2013). Morphadorner v2: A java library for the morphological adornment of english language texts. *Northwestern University, Evanston, IL.*

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Coster, W. and Kauchak, D. (2011). Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th ACL*, pages 665–669.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th WMT*, pages 85–91. Association for Computational Linguistics.

Devlin, S. and Tait, J. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Devlin, S. (1999). *Simplifying Natural Language for Aphasic Readers*. Ph.D. thesis, University of Sunderland.

Feblowitz, D. and Kauchak, D. (2013). Sentence simplification as tree transduction. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10.

Glavaš, G. and Štajner, S. (2015). Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd ACL*, pages 63–69.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36(2):193–202.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010). A

[4] https://github.com/ghpaetzold/massalign

semantic and syntactic text simplification tool for health content. In *Proceedings of the 2010 AMIA*, pages 366–70.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstration session*, Prague, Czech Republic.

Paetzold, G. H. and Specia, L. (2013). Text simplification as tree transduction. In *Proceedings of the 9th STIL*, pages 116–125.

Paetzold, G. H. and Specia, L. (2015). Lexenstein: A framework for lexical simplification. In *Proceedings of The 53rd ACL*, pages 85–90.

Paetzold, G. H. and Specia, L. (2016a). Benchmarking lexical simplification systems. In *Proceedings of the 10th LREC*.

Paetzold, G. H. and Specia, L. (2016b). Inferring psycholinguistic properties of words. In *Proceedings of the 2016 NAACL*, pages 435–440.

Paetzold, G. H. and Specia, L. (2016c). Unsupervised lexical simplification for non-native speakers. In *Proceedings of The 30th AAAI*, pages 3761–3767.

Paetzold, G. H. and Specia, L. (2016d). Vicinity-driven paragraph and sentence alignment for comparable corpora. *arXiv preprint arXiv:1612.04113*.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). Massalign: Alignment and annotation of comparable documents. In *Proceedings of the 8th IJCNLP*, pages 1–4.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*.

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

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the 2010 HLT*, pages 403–411.

Štajner, S., Franco-Salvador, M., Ponzetto, S. P., Rosso, P., and Stuckenschmidt, H. (2017). Sentence alignment methods for improving text simplification systems. In *Proceedings of ACL*, pages 97–102.

Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 EMNLP*, pages 409–420.

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. *TACL*, 4:401–415.

Yasseri, T., Kornai, A., and Kertész, J. (2012). A practical approach to language complexity: a wikipedia case study. *PloS one*, 7(11):e48386.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *NAACL*, pages 365–368, Los Angeles, California.

Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of EMNLP*, pages 595–605.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. *Computational Linguistics*, (August):1353–1361.

9. Language Resource References

Newsela. (2016). *Newsela Article Corpus*. Version: 2016-01-29, `https://newsela.com/data`.