

# Learning Simplifications for Specific Target Audiences

Carolina Scarton and Lucia Specia

Department of Computer Science, University of Sheffield  
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK  
{c.scarton, l.specia}@sheffield.ac.uk

## Abstract

Text simplification (TS) is a monolingual text-to-text transformation task where an original (complex) text is transformed into a target (simpler) text. Most recent work is based on sequence-to-sequence neural models similar to those used for machine translation (MT). Different from MT, TS data comprises more elaborate transformations, such as sentence splitting. It can also contain multiple simplifications of the same original text targeting different audiences, such as school grade levels. We explore these two features of TS to build models tailored for specific grade levels. Our approach uses a standard sequence-to-sequence architecture where the original sequence is annotated with information about the target audience and/or the (predicted) type of simplification operation. We show that it outperforms state-of-the-art TS approaches (up to 3 and 12 BLEU and SARI points, respectively), including when training data for the specific complex-simple combination of grade levels is not available, i.e. zero-shot learning.

## 1 Introduction

Text simplification (TS) is the task of modifying an original text into a simpler version of it. One of the main parameters for defining a suitable simplification is the target audience. Examples include elderly, children, cognitively impaired users, non-native speakers and low-literacy readers.

Traditionally, work on TS has been divided in lexical simplification (LS) and syntactic simplification (SS). LS (Paetzold, 2016) deals with the identification and replacement of complex words or phrases. SS (Siddharthan, 2011) performs

structural transformations such as changing a sentence from passive to active voice. However, most recent approaches learn transformations from corpora, addressing simplification at lexical and syntactic levels altogether. These include either learning tree-based transformations (Woodsend and Lapata, 2011; Paetzold and Specia, 2013) or using machine translation (MT)-based techniques (Zhu et al., 2010; Coster and Kauchak, 2011a; Wubben et al., 2012; Narayan and Gardent, 2014; Nisioi et al., 2017; Zhang and Lapata, 2017). This paper uses the latter type of technique, which treats TS as a monolingual MT task, where an original text is “translated” into its simplified version.

In order to build MT-based models, a parallel corpus of original texts with their simplified counterparts is needed. For English, two main such corpora are available: Wikipedia-Simple Wikipedia (W-SW) (Zhu et al., 2010) and the Newsela Article Corpus.<sup>1</sup> The former is a collection of original Wikipedia articles and their simplified versions created by volunteers. The latter consists of news articles professionally simplified for various specific audiences following the US school grade system. To build simplification models, the pairs of articles in these corpora have been aligned at the level of smaller units using standard algorithms (Coster and Kauchak, 2011b; Paetzold and Specia, 2016; Štajner et al., 2017). Based on the number of sentences involved in these alignments, one can categorise alignments into four types of coarse-grained simplification operations:

- Identical: an original sentence is aligned to itself, i.e. no simplification is performed.
- Elaboration: an original sentence is aligned to a single, rewritten simplified sentence.
- One-to-many: splitting – an original sentence is aligned to 2+ simplified sentences.

<sup>1</sup><https://newsela.com/data, v.2016-01-29>.

- Many-to-one: joining – 2+ original sentences are aligned to a single simplified sentence.

We hereafter refer to the unit of simplification, i.e. one or more original or simplified sentences, as *instances*.

The Newsela corpus is seen as having higher quality than W-SW because its simplifications are created by professionals, following well defined guidelines (Xu et al., 2015). It is also larger which is preferable for training corpus-based models. More interestingly, the Newsela corpus has a feature that has been ignored thus far: Each instance in the corpus was created for readers with a certain school grade level. Each original article has a label indicating its corresponding grade level (from 12 to 2), and may have various simplified versions, each for a different grade level. For example, a level 12 article may have simplified counterparts for levels 8 and 4. In other words, the corpus contains instances where the same input leads to different outputs. Disregarding this factor may lead to suboptimal models. To avoid this problem, previous work (Alva-Manchego et al., 2017; Zhang and Lapata, 2017; Scarton et al., 2018b) has used subsets of the corpus with only certain combinations of complex-simplified article pairs, e.g. adjacent or non-adjacent pairs. This however reduces the amount of data available for training.

We propose a way of making use of this information to build more informed TS models that are aware of different types of target audiences, while still making use of the full dataset for learning. Inspired by the work of Johnson et al. (2017) for MT, we add to each original instance an artificial token that represents the target grade level of that instance in order to guide a sequence-to-sequence attentional encoder-decoder neural approach (Bahdanau et al., 2015) (§2). In a similar vein, we also annotate the coarse-grained type of operation that should be performed to simplify the original instance, under the hypothesis that certain operations are more often used to simplify into certain grade levels. Deciding on the operation is an easier problem than performing the actual operation. We rely on both gold and predicted operation types.

Experiments with models built with these artificial tokens outperform state-of-the-art neural models for TS, with the best approach combining grade level and type of operation (§3). Interestingly, such an approach also enables zero-shot TS, where a simplification for a grade level pair unseen

at training time can still be generated during testing. We show that our zero-shot learning models perform virtually as well as our grade/operation-informed models (§4). To the best of our knowledge, this is the first work to build TS models for specific target audiences and to explore zero-shot learning for this application.

## 2 System architecture

Our approach follows that of Johnson et al. (2017), a multilingual MT approach that adds an artificial token to encode the target language to the beginning of each source sentence in the parallel corpus. With this modified version of the corpus, a single encoder-decoder architecture is used to deal with different language pairs. Based on the tokens, the source sentences are encoded differently according to the target language they have been paired with in the corpus. Such an approach enables zero-shot MT, where a model is able to provide translations for language pairs it has not seen at training time.

We apply three types of data manipulation, where artificial tokens are added to the beginning of original side of both training and test instances:

- **to-grade:** the token corresponds to the grade level of the target instance,
- **operation:** the token is one of the four possible coarse-grained operations that transforms the original into the simplified instance,
- **to-grade-operation:** concatenation of the two above tokens.

Different from the grade level, which can be available at test time simply by knowing the intended reader of the text, information about the operations to be performed, which we extracted from the parallel corpus, will not be available at test time. We use gold labels extracted from the parallel corpus for an oracle experiment but also use a classifier that predicts the operations for the test set based on those in the training data. We built a simple Naive Bayes classifier using the `scikit-learn` toolkit (Pedregosa et al., 2011) and nine features (Scarton et al., 2017):

- number of tokens / punctuation / content words / clauses,
- ratio of the number of verbs / nouns / adjectives / adverbs / connectives to the number of content words.

Table 1 shows examples of the tokens used when an original instance is marked to be simpli-

	to grade level 4	to grade level 2
to-grade	<4> dusty handprints stood out against the rust of the fence near Sasabe.	<2> dusty handprints stood out against the rust of the fence near Sasabe.
operation	<identical> dusty handprints stood out against the rust of the fence near Sasabe.	<elaboration> dusty handprints stood out against the rust of the fence near Sasabe.
to-grade-operation	<4-identical> dusty handprints stood out against the rust of the fence near Sasabe.	<2-elaboration> dusty handprints stood out against the rust of the fence near Sasabe.
reference	dusty handprints stood out against the rust of the fence near Sasabe.	dusty handprints could be seen on the fence near Sasabe.

Table 1: Examples of artificial tokens used.

fied to grade level 4 or grade level 2. Since the reference for grade level 4 is a copy of the original, the operation token for this case is <identical>. For level 2 the reference is a rewrite and, therefore, the operation token is <elaboration>.

We use OpenNMT<sup>2</sup> as our encoder-decoder architecture. Both encoder and decoder have two LSTM layers, hidden states of size 500 and dropout = 0.3. Global attention combined with input-feeding is used, as describe in (Luong et al., 2015). A model is trained for each dataset constructed with different artificial tokens for 13 epochs. The best model is selected according to perplexity on the development set. Figure 1 shows the architecture of the neural network, including attention and input-feeding. In this example, <token> represents the artificial token added to the pre-processed data.

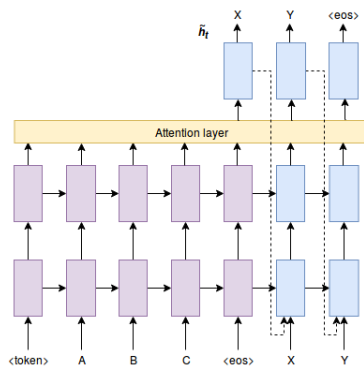


Figure 1: Neural model architecture.

We evaluate our models with BLEU<sup>3</sup> (Papineni et al., 2002) (a proxy for grammaticality assessment), SARI (Xu et al., 2016)<sup>4</sup> (a proxy for simplicity assessment) and Flesch Reading Ease<sup>5</sup> (a

<sup>2</sup>Torch version: <http://opennmt.net/OpenNMT/>

<sup>3</sup>The multi-blue.perl script from <https://github.com/moses-smt/mosesdecoder>

<sup>4</sup><https://github.com/cocoxu/simplification>

<sup>5</sup><https://github.com/mmautner/readability>

proxy for readability assessment). According to Xu et al. (2016), BLEU shows high correlation with human scores for grammaticality and meaning preservation, whilst SARI shows high correlation with human scores for simplicity. Although previous work have also relied on human judgements of grammaticality, meaning preservation and simplicity, in our case such a type of evaluation is infeasible: we would need to involve judges with specific grade levels or rely on professionals who are experts in grade level-specific simplification to make such assessments.

### 3 Reader-specific TS models

Our version of the Newsela corpus has 550,644 instance pairs (11M original tokens and 10M target tokens), which we randomly divided into training (440,516 instances: 80%), development (55,064 instances: 10%) and test (55,064 instances: 10%) sets. Instances were aligned using the method by Paetzold and Specia (2016). Xu et al. (2015) report over 56K original sentences and approximately 305K sentences including the original ones and all simplification types. Our number of instance pairs is higher because we allowed alignments from original to all simplified versions and among simplified versions. An original article 0 may be aligned to up to four simplified versions: 1, 2, 3 and 4. For each article, the alignments were extracted between 0- $\{1,2,3,4\}$ , 1- $\{2,3,4\}$ , 2- $\{3,4\}$  and 3-4, where available. Our corpus is also larger than the ones used in (Alva-Manchego et al., 2017; Scarton et al., 2018b) and (Zhang and Lapata, 2017). While the former use only adjacent levels (e.g. 0-1, 1-2) and the latter only non-adjacent levels (e.g. 0-2, 1-4), we make use of the full dataset.

As baseline we trained a model using OpenNMT and the same hyperparameters as described in §2 on the entire Newsela corpus but without artificial tokens (s2s model). The state-of-the-

art model is represented by NTS, which was also trained on the entire corpus using a similar OpenNMT architecture with the same hyperparameters but additional pre-trained word embeddings as described in Nisioi et al. (2017).<sup>6</sup>

As shown in Table 2 the NTS system performs slightly worse than the baseline system according to BLEU and SARI. Although concatenating global and local embeddings has led to improvements for the W-SW corpus in (Nisioi et al., 2017), this does not seem to be the case for the Newsela corpus. Our models outperform both the baseline and NTS systems by a large margin. Examples of outputs from all systems can be found in the Supplementary Material.

	BLEU $\uparrow$	SARI $\uparrow$	Flesch $\uparrow$
NTS	61.60	33.40	79.95
s2s	61.78	33.72	79.86
s2s+to-grade	<b>62.91</b>	<b>41.04</b>	82.91
s2s+operation (pred)	59.83	37.36	<b>84.96</b>
s2s+to-grade+operation (pred)	61.48	40.56	83.11
s2s+operation (gold)	63.24	41.81	84.47
s2s+to-grade+operation (gold)	<b>64.78</b>	<b>45.41</b>	<b>85.44</b>

Table 2: Results on the Newsela test set.

The best model is the one built with the <to-grade+operation> token with gold operations annotations (last row). The second best system uses the gold <operation> token only. Therefore, knowing the operation type to be performed for a given instance provides valuable information. Even though the models with predicted operations (‘pred’ in Table 2) still outperform the baseline, they lag behind their counterparts built using gold operations. The main reason for that is the very simplistic classifier we used (average accuracy = 0.51, calculated using 10-fold cross-validation). In summary, s2s+to-grade is the best performing model in a real world scenario, given the low performance of ‘pred’ systems. A more informed classifier should lead to better results, but this left for future work; our goal was to show the potential of this information.

The improvements in SARI are substantial: 7 points over the baseline even with the predicted operations. However, SARI aims to measure simplicity in general (not for specific grade levels). Since human evaluation of the targeted simplification performed by our models is not feasible, we can only approximate the usefulness of our models by using readability metrics such

<sup>6</sup>Equivalent to their best performing “NTS-w2v” version.

as the Flesch-Kincaid Grade Level. This metric maps a text into a US grade level, which is the same grading provided in the Newsela corpus and, therefore, relevant for our study. Table 3 shows the Flesch-Kincaid results for the test set divided into the appropriate grade levels considering the outputs of s2s, s2s+to-grade and s2s+to-grade+operation (gold) models. Simplifications generated by s2s+to-grade and s2s+to-grade+operation are scored consistently closer to the appropriate grade, which does not happen with s2s.

	s2s	+to-grade	+to-grade+operation (gold)
<10>	9.23	11.90	9.93
<9>	8.85	9.82	8.57
<8>	7.47	8.46	7.58
<7>	7.81	7.79	6.89
<6>	6.99	6.48	5.57
<5>	5.58	5.05	4.49
<4>	5.90	3.85	3.28
<3>	5.15	2.44	1.88
<2>	3.94	1.57	1.00
MAE	1.09	<b>0.63</b>	<b>0.63</b>

Table 3: Flesch-Kincaid scores for instances of each grade level simplified using s2s, s2s+to-grade and s2s+to-grade+operation (gold) models.

The last row of Table 3 shows the Mean Absolute Error (MAE) considering the Flesch-Kincaid Grade Level scores for the system outputs as the hypothesis and the expected grade level as the gold scores. Our s2s+to-grade and s2s+to-grade+operation (gold) models show lower error scores than the baseline system, which supports our hypothesis that such models produce more adequate outputs for targeted grade levels.

### 3.1 Usefulness of the s2s+to-grade model

The main advantage of s2s+to-grade is that a user can inform their grade level and retrieve a personalised simplification. Table 4 shows an example with different simplifications for an out-of-domain instance from the SimPA corpus (Scarton et al., 2018a). The same instance was given as input to the s2s+to-grade model with different artificial tokens according to the grade level that we want to achieve. The s2s system (second row) repeats the original instance (first row). Conversely, our s2s+to-grade model is capable of distinguishing among different levels and produces personalised simplifications for each grade level.

original	We want to reassure you that we take fire safety very seriously and we are doing everything we can to make sure our residents are safe.
s2s	We want to reassure you that we take fire safety very seriously and we are doing everything we can to make sure our residents are safe.
<10>	We want to reassure you that we take fire safety very seriously and we are doing everything we can to make sure our residents are safe.
<9>	We want to reassure you that we take fire safety very seriously and we are doing everything we can to make sure our residents are safe.
<8>	We want to reassure you that we take fire safety very seriously and we are doing everything we can to make sure our residents are safe.
<7>	We want to reassure you that we take fire safety very seriously and we are doing everything we can to make sure our residents are safe.
<6>	We want to reassure you that we take fire safety very seriously. <b>We</b> are doing everything we can to make sure our residents are safe.
<5>	We want to reassure you that we take fire safety very seriously. <b>We</b> are doing everything we can to make sure our residents are safe.
<4>	We want to <b>make sure</b> we take fire safety very seriously. <b>We</b> are doing everything we can to make sure our <b>people</b> are safe.
<3>	We want to <b>make sure</b> people take fire safety very seriously. <b>We</b> are doing everything we can to make sure our <b>people</b> are safe.
<2>	We want to <b>make sure</b> people take fire safety very seriously. <b>We</b> are doing everything we can to make sure <b>people</b> are safe.

Table 4: Examples of s2s+to-grade outputs when an original instance is simplified into different levels.

## 4 Zero-shot TS models

To show that zero-shot TS is possible, we build models on training data without instances of a certain grade level pair and test them on instances of that grade level pair. Consider the grade level pair  $\langle g_o, g_t \rangle$ , where  $g_o$  is the grade level of an original instance  $o$ ,  $g_t$  of a target instance  $t$ , and  $t$  is aligned to  $o$ . We test if our “s2s+to-grade” model can generalise for instances of  $\langle \hat{g}_o, \hat{g}_t \rangle$  that have not been seen at training time.

Due to space restrictions, we only show results for three representative grade level pairs. These pairs have a large enough number of training and test instances and cover levels that are closer or further apart from each other. In addition, after removing them the training corpus still has enough instances of the  $\hat{g}_t$  as target grade level. Instances of the target but not the original level (or of the target language in MT) must exist for zero-shot to be possible. The distributions of the selected grade level pairs is shown in Table 5.

	# training	# test	# of remaining $\hat{g}_t$
< 12, 7 >	30,246	3,825	34,545
< 12, 4 >	22,709	2,867	104,833
< 6, 5 >	18,122	2,239	79,546

Table 5: Zero-shot data distribution.

In Table 6, the s2s and s2s+to-grade models are the same as in Section 3, i.e. trained with the entire dataset without artificial tokens (s2s) or with artificial tokens (s2s+to-grade). The zero-shot models (s2s+to-grade+zs) are trained with <to-grade> data, but after removing instances of the grade level pair  $\langle \hat{g}_o, \hat{g}_t \rangle$  under investigation, i.e. on a smaller dataset. For < 12, 7 > and < 12, 4 >, the zero-shot models outperform the baseline according to all metrics. In terms of SARI, for < 12, 7 > the zero-shot model is only marginally worse than the s2s+to-grade model. Conversely, s2s+to-grade+zs outperforms s2s+to-grade for <

12, 4 >, which is an impressive result. Finally, for < 6, 5 > all three models perform similarly. This may be explained by the proximity of  $\hat{g}_o$  and  $\hat{g}_t$ , which means that instances must be considerably close to each other and therefore simplifications will be minor and have little impact in the scores.

	BLEU $\uparrow$	SARI $\uparrow$	Flesch $\uparrow$
< 12, 7 >			
s2s	63.02	38.43	73.83
s2s+to-grade	64.16	<b>40.61</b>	<b>74.43</b>
s2s+to-grade+zs	<b>64.50</b>	39.83	74.09
< 12, 4 >			
s2s	44.56	37.56	79.50
s2s+to-grade	49.43	50.76	91.04
s2s+to-grade+zs	<b>50.18</b>	<b>50.85</b>	<b>91.08</b>
< 6, 5 >			
s2s	<b>69.71</b>	<b>26.47</b>	84.74
s2s+to-grade	69.39	26.32	<b>87.07</b>
s2s+to-grade+zs	68.78	26.23	86.80

Table 6: Results of zero-shot experiments for TS.

## 5 Conclusions

We have presented an approach for TS that benefits from corpora built for various target audiences and allows building better models than general-purpose ones. We have also shown that zero-shot learning is possible for TS, where instances of the original-target audience do not exist. As future work we intend to investigate (i) better classifiers to predict operation types and (ii) multi-task learning as an alternative way of building a single TS model for various specific target audiences. We also plan to run experiments with the W-SW corpus and using an improved classifier to train models with information on operations.

## Acknowledgements

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

## References

- Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 295–305. <http://www.aclweb.org/anthology/I17-1030>.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, CA, ICLR '15. <https://arxiv.org/pdf/1409.0473.pdf>.
- William Coster and David Kauchak. 2011a. Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. ACL, Portland, Oregon, MTTG '11, pages 1–9. <http://dl.acm.org/citation.cfm?id=2107679.2107680>.
- William Coster and David Kauchak. 2011b. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 665–669. <http://dl.acm.org/citation.cfm?id=2002736.2002865>.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics* 5:339–351. <http://aclweb.org/anthology/Q17-1024>.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. <http://aclweb.org/anthology/D15-1166>.
- Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 435–445. <http://www.aclweb.org/anthology/P14-1041>.
- Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 85–91. <http://aclweb.org/anthology/P17-2014>.
- Gustavo Paetzold and Lucia Specia. 2013. Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. Fortaleza, Brazil, STIL, pages 116–125. <http://aclweb.org/anthology/W13-4813>.
- Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield, Sheffield, UK. <http://etheses.whiterose.ac.uk/15332/>.
- Gustavo Henrique Paetzold and Lucia Specia. 2016. Vicinity-driven paragraph and sentence alignment for comparable corpora. *CoRR* abs/1612.04113. <http://arxiv.org/abs/1612.04113>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL, Philadelphia, Pennsylvania, ACL '02, pages 311–318. <https://doi.org/10.3115/1073083.1073135>.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.
- Carolina Scarton, Gustavo Henrique Paetzold, and Lucia Specia. 2018a. SimPA: A Sentence-Level Simplification Corpus for the Public Administration Domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Miyazaki, Japan, pages 4333–4338. <http://www.lrec-conf.org/proceedings/lrec2018/pdf/542.pdf>.
- Carolina Scarton, Gustavo Henrique Paetzold, and Lucia Specia. 2018b. Text Simplification from Professionally Produced Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Miyazaki, Japan, pages 3504–3510. <http://www.lrec-conf.org/proceedings/lrec2018/pdf/1063.pdf>.
- Carolina Scarton, Alessio Palmero Arosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A Multilingual Syntactic Simplification Tool. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (System Demonstrations)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 25–28. <http://aclweb.org/anthology/I17-3007>.

- Advait Siddharthan. 2011. Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Stroudsburg, PA, USA, ENLG '11, pages 2–11. <http://dl.acm.org/citation.cfm?id=2187681.2187684>.
- Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 97–102. <http://aclweb.org/anthology/P17-2016>.
- Kristian Woodsend and Mirella Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 409–420. <http://www.aclweb.org/anthology/D11-1038>.
- Sander Wubben, Antal van den Bosch, and Emiel Kraemer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 1015–1024. <http://www.aclweb.org/anthology/P12-1107>.
- Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics* 3:283–297. <http://www.aclweb.org/anthology/Q15-1021>.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* 4:401–415. <http://www.aclweb.org/anthology/Q16-1029>.
- Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 595–605. <https://www.aclweb.org/anthology/D17-1063>.
- Zheming Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '10, pages 1353–1361. <http://dl.acm.org/citation.cfm?id=1873781.1873933>.