SimPA: A Sentence-Level Simplification Corpus for the Public Administration Domain

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

Abstract

We present a sentence-level simplification corpus with content from the Public Administration (PA) domain. The corpus contains $1,100$ original sentences with manual simplifications collected through a two-stage process. Firstly, annotators were asked to simplify only words and phrases (lexical simplification). Each sentence was simplified by three annotators. Secondly, one lexically simplified version of each original sentence was further simplified at the syntactic level. In its current version there are $3,300$ lexically simplified sentences plus $1,100$ syntactically simplified sentences. The corpus will be used for evaluation of text simplification approaches in the scope of the EU H2020 SIMPATICO project – which focuses on accessibility of e-services in the PA domain – and beyond. The main advantage of this corpus is that lexical and syntactic simplifications can be analysed and used in isolation. The lexically simplified corpus is also multi-reference (three different simplifications per original sentence). This is an ongoing effort and our final aim is to collect manual simplifications for the entire set of original sentences, with over 10K sentences.

Keywords: text simplification, simplification corpora, public administration

1. Introduction

Text simplification (TS) is the task of reducing lexical and/or structural complexity of texts (Siddharthan, 2004). It is common to divide this task in two: lexical simplification (LS) and syntactic simplification (SS). LS deals with the identification and replacement of difficult words or phrases, while SS focuses on making complex syntactic structures simpler, e.g. by changing passive into active voice or splitting a sentence with coordination in two sentences.

LS has been widely explored in recent years. Work include simple word frequency-based approaches (Carroll et al., 1998; Carroll et al., 1999; Biran et al., 2011), unsupervised approaches that use word embeddings (Glavaš and Štajner, 2015; Paetzold and Specia, 2016c), and supervised approaches (Horn et al., 2014; Paetzold and Specia, 2017). For SS, some approaches apply hand-crafted rules (Siddharthan, 2011; Candido Jr. et al., 2009; Bott et al., 2012; Brouwers et al., 2014; Barlacchi and Tonelli, 2013; Scarton et al., 2017), while others use parallel data to learn simplification operations (Woodsend and Lapata, 2011; Paetzold and Specia, 2013; Siddharthan and Angrosh, 2014; Zhu et al., 2010; Coster and Kauchak, 2011; Wubben et al., 2012; Narayan and Gardent, 2014; Xu et al., 2016; Zhang and Lapata, 2017; Nisioi et al., 2017). Corpus-based approaches tend to learn both LS and SS transformations jointly.

For English, two main parallel corpora exist: Simple Wikipedia (Zhu et al., 2010) and Newsela (Newsela, 2016), the latter with simplifications performed by professionals. However, such corpora do not distinguish among different types of simplification (i.e. lexical from syntactic transformations). Moreover, they cover general domain texts, and therefore may not be sufficient to model operations for specific domains.

In terms of specific corpora for TS evaluation, only a few exist, mostly for English LS (Horn et al., 2014; De Belder and Moens, 2012; Paetzold and Specia, 2016c; Paetzold and Specia, 2016a), all of which are composed of a sentence, a target complex word, and candidate substitutions ranked by simplicity. Although these have been used in many papers, their sentences are extracted either from Wikipedia or news articles, making them also general domain. The candidate substitutions in most of these datasets were suggested and ranked by native English speakers, which means that they do not necessarily capture the needs of specific audiences, such as non-native speakers.

For corpus-based TS – including both LS and SS – Zhu et al. (2010) released a test set of 100 original sentences and 131 simplified sentences, a subset from the Simple Wikipedia corpus. Xu et al. (2016) released a test set of 350 sentences with nine simplifications each sentence. They also used a subset of Simple Wikipedia, but containing only 1-to-1 aligned sentences, i.e. they disregarded sentence splitting, a very common operation where one sentence is broken into two or more.

These datasets are therefore either very small, have only one reference simplification, or do not cover all types of simplification. Simplification is a complex process and often more than one possible way of modifying a sentence is possible and acceptable. In addition, existing datasets are not suitable for approaches targeting a specific domain or user type. In the SIMPATICO project[1] we address the simplification of Public Administration (PA) content, such as websites that describe services, citizen rights and duties. Among the target audiences are non-native speakers of English. Our ultimate goal is to be able to provide personalised lexical and syntactic simplifications for each this target audience. In order to tackle the lack of evaluation data, we introduce SimPA: an English sentence-level TS evaluation corpus for the PA domain.

The SimPA corpus is under construction. The ultimate goal

[1] https://www.simpatico-project.eu

is to have at least one simplification for each original sentence in an entire corpus of $10,708$ original sentences that cover different types of PA-related content. he resulting corpus could be used not only for test, but also for development or fine-tuning of corpus-based approaches trained on corpora from other domains, such as the Wikipedia corpus. The current release contains $1,100$ PA domain sentences with three lexically simplified versions each, and one version further annotated with syntactic simplification. Lexical and syntactic simplification were done separately in order to build a resource with information about both tasks independently. By isolating such operations, our corpus can be directly used to evaluate specialised systems, that only perform either lexical or syntactic simplification. Nonetheless, this characteristic of the corpus should not impact its use for the evaluation of systems that perform both operations together. An user of the corpus just needs to be aware of which version of the corpus to use for each purpose. In addition to evaluation, this corpus will help with the analysis and profiling of domain-specific simplifications for the better design of simplification systems that better suit the purposes of our target audience.

Even in its current status, SimPA is the largest and most varied dataset of its kind, with multiple references for the LS version. SimPA is freely available under a Creative Commons Licence.[2]

2. Corpus Creation

In order to create SimPA, we first collected sentences from the Sheffield City Council (SCC), website[3] which is one of the partner PAs in the SIMPATICO project. Our crawler visited over 9K links, resulting in around 14K sentences. We then filtered the sentences to eliminate those without verbs, such as titles, menu items, and incomplete sentences. We also removed repeated sentences. This resulted in $10,708$ sentences, which we refer to as the 10K corpus.

In order to start with potentially more challenging sentences, maximising the opportunities that annotators would have to perform both lexical and syntactic simplifications, we sorted the 10K sentences according to their length (longer sentences first). The 5K longest sentences were then shuffled and $1,100$ sentences were selected to be annotated. Table 1 shows some statistics about the 10K, 5K and 1.1K sets.

	# tokens	# sentences	tokens per sentence
10K	249,954	10,708	23.34
5K	153,680	5,000	30.74
1.1K	33,492	1,100	30.45

Table 1: Statistics of the PA corpus.

2.1. Gathering Lexical Simplifications

The first step was to obtain lexical simplifications for our PA sentences. We define lexical simplifications as any word and phrase replacements that make the sentence more easily understood. In other words, a lexical simplification modifies the sentence locally, without altering sentence grammaticality or compromising meaning. We have collected three lexically simplified versions of each of our 1,100 PA subset sentences. This allows the creation of a more reliable evaluation dataset and the analysis of how people with different profiles attempt to minimise the lexical complexity of a sentence.

The data annotation was conducted using volunteers, all students and academic staff from universities, who considered themselves fluent speakers of English (native and non-native). Annotators were anonymised but asked to inform some demographic information (Section 3.1.).

Each annotator received 20 sentences to simplify on a online form created using the Google Forms platform.[4] They were given the following guidelines: "*Replace any words and phrases that you find complex in each sentence. Simpler words or phrases are those that you think can be more easily understood by readers, especially non-native speakers of English. Do not change the sentence in any other way.*"

In order to encourage annotators to produce reliable annotations, they could enter a £50 prize draft by providing their email. Inspecting the annotations, we found and replaced 229 spurious simplified sentences, most of which were produced by untrustworthy annotators.

The current version of SimPA has a total of $3,300$ lexically simplified PA sentences (3 versions of $1,100$ PA sentences). Table 2 shows some examples, where words in bold are the changes made by annotators (red is used to mark the original words/phrases and blue is used to mark the simplifications). They include replacements of single words (such as "experiencing" vs. "hearing"), as well as phrases (such as "a period of public consultation" vs. "consulting the public").

2.2. Gathering Syntactic Simplifications

The next step was to obtain syntactic simplifications for the lexically simplified sentences. These are any transformations that alter the syntactic structure of the sentence, such as splitting, passive-to-active voice transformation, anaphoric resolution, information reordering, etc. For this stage, thus far we collected only one simplification per sentence.

We selected one lexically simplified version of each original $1,100$ sentences from our previous annotation step. For cases with more than one distinct simplification, we randomly selected one. In this way, we always selected a lexically simplified sentence, if available. Although the sentence selection process could have involved some kind of readability assessment, we opted for the random approach, since we assumed all three simplifications are valid and correct. $1,079$ sentences from the final set featured at least one lexical simplification, only 21 did not.

Much like in the previous step, the annotators were students and academic staff from universities, and we also used the Google Forms platform. The same demographic data was collected and the number of sentences to simplify given to

[2] https://github.com/SIMPATICOProject/simpa

[3] https://www.sheffield.gov.uk

[4] https://www.google.co.uk/forms/about

Original:	If you're **experiencing** excessive noise form a **commercial property** (such as a pub, club, restaurant or cafe) you can report it to us.
Simplified:	If you're **hearing** excessive noise from a **business** (such as a pub, club, restaurant or cafe) you can report it to us.
Original:	After **a period of public consultation**, we adopted the **appraisal** and accompanying management **proposals** on 23rd October 2007.
Simplified:	After **consulting the public**, we adopted the **review** and accompanying management **plans** on 23rd October 2007.
Original:	In my personal opinion, I **don't** think the Housing Service is **given** enough **credit** for the amount of work they do for their tenants.
Simplified:	In my own opinion, I **do not** think the Housing Service is **praised** enough for the amount of work they do for their tenants.
Original:	**Where** agreement cannot be reached at the **compliance** stage and **liability** orders are moved to the enforcement stage a further £235 will be added to the **debt** owed.
Simplified:	**If** agreement cannot be reached at the **verification** stage and **responsibility** orders are moved to the enforcement stage a further £235 will be added to the **bill** owed.
Original:	The review considered **the suitability of** the technical standards being used by Local Planning Authorities and proposed a **radical** reduction in the number of **eligible** standards.
Simplified:	The review considered **how good** the technical standards being used by Local Planning Authorities **were** and proposed a **big** reduction in the number of **acceptable** standards.

Table 2: Example of manual LS

Original:	According to law, a successful challenge would result in a acquisition exercise **in which** the challenger could take part along with other interested organisations.
Simplified:	According to law, a successful challenge would result in an acquisition exercise**. In this exercise** the challenger could take part, as well as other interested organisations.
Original:	Within the 28 days application period we will talk with South Yorkshire Police for any comments and take into account any rules.
Simplified:	We will talk with the South Yorkshire Police for any comments**. This will be done** within the 28 days application period**. We will then** take into account any rules.
Original:	The number of dogs and cats that may be accommodated **will be specified** on the licence along with any other specific conditions.
Simplified:	The license **will have** the number of dogs and cats that can be accommodated along with any other specific conditions.
Original:	If required, the developer has to get together an ES describing the likely effects of the development on the environment and suggested measures to reduce problems.
Simplified:	If required, the developer has to get together an ES**. This describes the possible** effects of the development on the environment**. It also includes** suggested measures to reduce problems.
Original:	**If** you can pay the full costs of your care and support without any financial help from us, **then** you are considered as a self-funder.
Simplified:	You are a self-funder **if** you can pay for your own care and support without receiving financial help.

Table 3: Example of manual SS

each annotator was also 20. They were provided with the following guidelines: "*Apply simplification operations to each sentence. You can split it, rewrite it, or reorder its information. Please avoid adding extra information or deleting parts of the sentence (unless it is extremely necessary). The goal is for the simplified sentence to have the same meaning as the original sentence.*". Annotators could again enrol in a £50 prize draft upon completion of the task.

Examples of annotations are shown in Table 3. The first line shows a case where a relative clause was split into two sentences. The original sentence in the second line was split into three sentences where two conjoint clauses (one temporal "within 28 days" and one of addition "and") were split. In the third line, the original sentence was changed from the passive into active voice. The fourth line also contains examples of splitting, where the original sentences was split into three. Finally, the fifth sentence was reordered (the "if" clause changed places with the "then" clause).

3. Data Analysis

3.1. Annotators

Since our aim is to create models capable of generating personalised simplifications, the demographic information we collected becomes very important. We can, for instance, identify which words or syntactic structures were modified by readers with a certain native language, or who are from a given country, or have the same proficiency with English. While we acknowledge that our sample is rather small to

draw conclusions, we can use the data to already provide insights on the perception of different readers when it comes to text complexity. The demographic questions were:

- Age;
- Country of birth;
- Native language (with the option to include up to three languages);
- Educational level;
- Proficiency in English (following the CEFR[5] scale);
- Familiarity with the PA e-services; and
- Occupation.

Although all volunteers are part of the academic environment, they still have very diverse backgrounds and profiles. In addition, volunteers included undergraduate and postgraduate students, and member of staff with a variety of educational levels. Finally, the volunteers were from different faculties of universities, including humanities, sciences, engineering, medical school, among others.

176 volunteers participated in the first annotation stage (lexical simplification). Figure 1 illustrates the distributions for each demographic aspect.[6] The number of volunteers that have more than one native language is 30.

For the second stage (syntactic simplification), 85 volunteers participated in the task. Figure 2 shows the distribution of each demographic aspect for all the annotators in the second phase. 11 volunteers reported having more than one native language.

For both stages, the majority of the volunteers are British and have English as their native language (which is expected given the survey distribution channels used). Brazil is the country with the second highest proportion of volunteers in both stages. The second most common native language was Portuguese, followed by Other/Unlisted and Chinese, respectively.

In general, the majority of the volunteers are between 18 and 30 years old. In the first stage, we had more undergraduate students, whilst in the second stage the majority were postgraduate students. In both experiments, the large majority of the volunteers are either C1 or C2 in the CEFR scale. This is also expected as we requested fluent speakers of English. Finally, more than half of our volunteers had no experience with PA e-services.

3.2. Simplification Data

The lexical simplifications for the $1,100$ sentences have $33,301$ tokens in total, which is 191 tokens less than the original sentences. This indicates that the majority of the changes were at word level and only a few phrases were simplified into single words. The syntactic simplifications have $32,219$ tokens, which is $1,082$ tokens less than their lexically simplified versions. This is expected, since some syntactic transformations tend to lead to the removal of unimportant words.

[5] http://www.coe.int/en/web/common-european-framework-reference-languages

[6] We omitted the data for occupation as it was very sparse.

Table 4 shows readability scores calculated for the original, lexically simplified and syntactically simplified sentences in SimPA. The basic counts of syllables per words (tokens except punctuation), words per sentence and content words (nouns, adjectives, verbs and adverbs) per sentence show that the LS version has less words than the original and that such words are shorter.

	Original	LS	SS
Syllables per word	1.91	1.85	1.86
Words per sentence	24.04	23.91	22.76
CW per sentence	16.15	15.90	11.90
Age of acquisition ↓	316.58	306.26	299.29
Familiarity ↑	439.54	445.32	440.36
Imageability ↑	315.37	317.64	314.86
Concreteness* ↑	298.90	299.83	297.12
Flesch Reading Ease ↑	44.8	48.6	100

Table 4: Readability metrics comparing the original and two simplified corpus. * indicates no statistically significant difference according to t-test with $p < 0.05$

We also evaluated the data using four psycholinguistics metrics: age of acquisition, familiarity, imageability and concreteness, that were extracted using the approach proposed by Paetzold and Specia (2016b). Age of acquisition is significantly smaller for the lexically simplified sentences, meaning that the words used are usually learned at a younger age. Familiarity and imageability are significantly higher for them also, which suggests that the words and phrases used to replace segments are simpler than the original.

The ratio of syllables per word is almost the same for lexically and syntactically simplified sentences. The number of words and content words per sentence is smaller in syntactic than in lexical simplifications, since volunteers shortened the sentences during syntactic simplification. The age of acquisition score is significantly smaller for syntactic simplifications than for lexical ones, which would suggest that a lot of syntactic simplifications also encompass word replacements. However, the values for familiarity, imageability and concreteness are smaller for syntactic simplifications, which contradicts this hypothesis.

According to Flesch, lexical simplifications were shown not to greatly affect the readability of sentences, increasing their scores by no more than 3.8 points. Although the difference is small, it is statistically significant (t-test with $p < 0.05$). Syntactic simplifications, on the other hand, achieved a much more impressive gain of 55.2 in Flesch, more than doubling the original sentences' readability.

4. Discussion

We presented the first version of SimPA: a dataset for the analysis and evaluation of simplifications for content from the public administration domain. The dataset is currently composed of three lexically simplified versions of $1,100$ PA sentences each, as well as one syntactically simplified version of $1,100$ lexically simplified sentences.

Our analyses reveal that SimPA contains simplifications produced by people with various backgrounds, meaning that it can be used to evaluate the performance of simplifiers

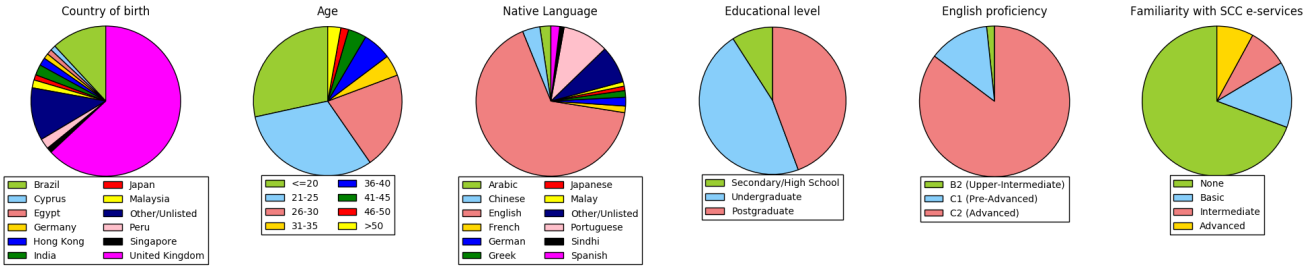

Figure 1: Demographic data of the lexical simplification experiment.

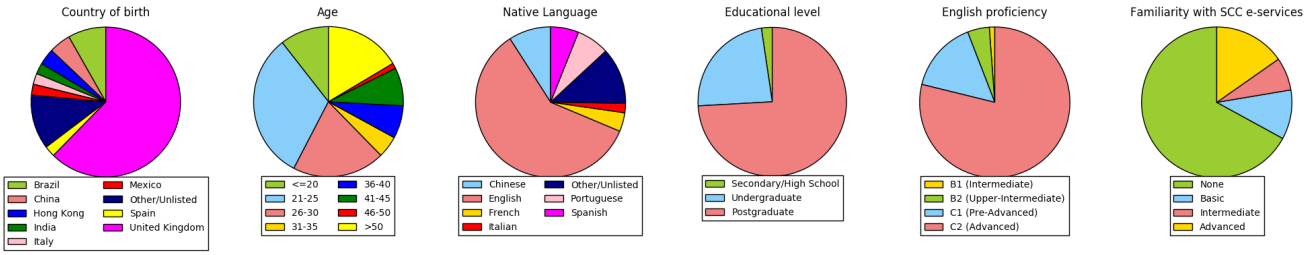

Figure 2: Demographic data of the syntactic simplification experiment.

that produce personalised output. By looking at psycholinguistic properties we found that the lexical simplifications feature many word and phrase replacements that reduce the overall age of acquisition, as well as increase the overall familiarity, imageability and concreteness of the original sentence. We also found that syntactic simplifications lead to sentences that are more than twice as readable as both the original and lexically simplified versions.

As future work, we intend to gather simplifications for the remaining of our corpus (9,608 sentences), in order to build a larger dataset for development or fine tuning purposes. For that, to make the process more efficient we will gather both lexical and syntactic simplifications in a single step (instead of splitting the process in lexical and syntactic simplification stages).

Acknowledgements

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

5. Bibliographical References

Barlacchi, G. and Tonelli, S. (2013). Ernesta: A sentence simplification tool for children's stories in italian. In *Proceedings of the 14th CICLing*, pages 476–487.

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., Saggion, H., and Mille, S. (2012). Text Simplification Tools for Spanish. In *Proceedings of the 8th LREC*, pages 1665–1671.

Brouwers, L., Bernhard, D., Ligozat, A.-L., and François, T. (2014). Syntactic Simplification for French. In *Proceedings of the 3rd PITR*, pages 47–56.

Candido Jr., A., Maziero, E., Gasperin, C., Pardo, T. A. S., Specia, L., and Aluísio, S. M. (2009). Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In *Proceedings of the 4th NAACL HLT BEA*, pages 34–42.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of 1998 AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 9th EACL*, pages 269–270.

Coster, W. and Kauchak, D. (2011). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the MTTG 2011*, pages 1–9.

De Belder, J. and Moens, M.-F. (2012). A dataset for the evaluation of lexical simplification. In *Computational Linguistics and Intelligent Text Processing*, pages 426–437.

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

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

Narayan, S. and Gardent, C. (2014). Hybrid Simplification using Deep Semantics and Machine Translation. In *Proceedings of the 52nd ACL*, pages 435–445.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *ACL (2)*. The Association for Computational Linguistics.

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. (2016a). Benchmarking Lexical Simplification Systems. In *Proceedings of the 10th LREC*, pages 3074–3080.

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. and Specia, L. (2017). Lexical simplification with neural ranking. In *Proceedings of the 15th EACL*, pages 34–40.

Scarton, C., Specia, L., Palmero Aprosio, A., Tonelli, S., and Martín Wanton, T. (2017). MUSST: A Multilingual Syntactic Simplification Tool. In *Proceedings of IJCNLP 2017*, pages 25–28, Taipei, Taiwan.

Siddharthan, A. and Angrosh, M. A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th EACL*, pages 722–731.

Siddharthan, A. (2004). *Syntactic simplification and text cohesion*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

Siddharthan, A. (2011). Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. In *Proceedings of the 13th ENLG*, pages 2–11.

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

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th ACL*, pages 1015–1024.

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

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. In *Proceedings of the 23rd COLING*, pages 1353–1361.

6. Language Resource References

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

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd COLING*, pages 1353–1361.