Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies

Sanja Štajner[1] and **Iacer Calixto**[2] and **Horacio Saggion**[3]
[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
SanjaStajner@wlv.ac.uk
[2]ADAPT Centre, Dublin City University, School of Computing, Ireland
icalixto@computing.dcu.ie
[3]TALN Research Group, Universitat Pompeu Fabra, Spain
horacio.saggion@upf.edu

Abstract

In this paper, we explore statistical machine translation (SMT) approaches to automatic text simplification (ATS) for Spanish. First, we compare the performances of the standard phrase-based (PB) and hierarchical (HIERO) SMT models in this specific task. In both cases, we build two models, one using the TS corpus with "light" simplifications and the other using the TS corpus with "heavy" simplifications. Next, we compare the two best systems with the state-of-the-art text simplification system for Spanish (Simplext). Our results, based on an extensive human evaluation, show that the SMT-based systems perform equally as well as, or better than, Simplext, despite the very small datasets used for training and tuning.

1 Introduction

The goal of automatic text simplification (ATS) is to transform lexically and syntactically complex texts or sentences into their simpler variants which can be more easily understood by non-native speakers, children, and people with various language or learning impairments (e.g. people with autism, dyslexia, or intellectual disabilities). Due to the scarcity and limited sizes of parallel corpora of original and manually simplified sentences, the state-of-the-art ATS systems are still predominantly rule-based for many languages, e.g. Spanish (Drndarević et al., 2013), Basque (Aranzabe et al., 2013), and French (Brouwers et al., 2014).

Recently, several studies proposed applying the standard PB-SMT model to the text simplification task for Brazilian Portuguese (Specia, 2010), English (Coster and Kauchak, 2011), and Spanish (Štajner, 2014). None of those studies, however, performed a thorough human evaluation of the systems or directly compared their systems to the existing rule-based ATS systems for those languages. The reported automatic evaluation (using BLEU score) gives us no insights on the correctness and usefulness of those systems and how well they perform in comparison to the state-of-the-art rule-based ATS systems.

In this paper, we address the problem of ATS for Spanish, investigating the possibility of applying the standard phrase-based (PB) and hierarchical (HIERO) SMT models to the only two currently-known text simplification (TS) parallel corpora for Spanish. We perform an extensive human evaluation of the generated output which allows us to compare the systems directly. Additionally, we compare our two best systems with Simplext, the state-of-the-art text simplification system for Spanish (Saggion et al., 2015).

Our experiments make several contributions to the field of automatic text simplification by exploring the following important questions:

1. How well can PB-SMT and HIERO models perform if built using very small parallel TS corpora?
2. Do the results obtained using standard PB-SMT models differ significantly from the ones obtained using the HIERO models?
3. How do the SMT-based models for ATS perform in comparison with the state-of-the-art ATS system for Spanish?

Proceedings of Recent Advances in Natural Language Processing, pages 618–626,
Hissar, Bulgaria, Sep 7–9 2015.

To the best of our knowledge, this is the first study (for any language) which applies a HIERO model to text simplification, and the first study which directly compares performances of the SMT-based models with a state-of-the-art ATS system.

2 Related Work

With the emergence of the Simple English Wikipedia[1], which together with the "original" English Wikipedia offered a large comparable text simplification (TS) corpus (137,000 sentence pairs), the focus of the ATS for English was shifted towards data-driven approaches. Most of them applied various SMT techniques, either phrase-based (Coster and Kauchak, 2011; Wubben et al., 2012), or syntax-based (Zhu et al., 2010; Woodsend and Lapata, 2011). In other languages, TS corpora either do not exist or they are very limited in size (only up to 1,000 sentence pairs). The only known exception to this is the case of Brazilian Portuguese for which there is a parallel TS corpus with 4,483 sentence pairs, built under the PorSimples project (Aluísio and Gasperin, 2010), aimed at simplifying texts for low literacy readers. This corpus has been used to train the standard PB-SMT model for ATS (Specia, 2010), and the reported results were promising (BLEU = 60.75) despite the small size of the dataset. The recent attempt at using the standard PB-SMT models for ATS for Spanish on two TS corpora of limited size (850 sentence pairs each) indicated that: (1) the level of simplification present in the datasets ("heavy" or "light") significantly influences the results, and (2) the model built using the "light" corpora can still learn some useful simplifications despite the very small size of the dataset (Štajner, 2014).

2.1 SMT for Low-Resourced Languages

The main problems in SMT applied to low-resourced languages ("simplified" Spanish can be seen as such) are the accuracy and coverage (Irvine and Callison-Burch, 2013). The first problem is the result of the fact that the model does not have enough data to estimate good probabilities over the possible translations and therefore ensure correctness of the translation pairs. The second problem occurs when the model and its word coverage are small, which leads to a high number of out-of-vocabulary (OOV) words. Words which the model have not encountered during the training phase cannot be correctly dealt with during the test phase.

2.2 Monolingual SMT

When monolingual SMT is used for text simplification, the problem of coverage is not so much of an issue as it is in cross-lingual SMT. In our case, the source language is the "regular" Spanish, and the target language is the "simplified" Spanish. Therefore, if a word in the source language is not found in the translation table – and is, therefore, an OOV word – it will be left untranslated. This might impact the overall simplicity of the output (in the case that the OOV word was complex), but it will not necessarily deteriorate the grammaticality and meaning preservation of the output sentence (as would be the case in cross-lingual SMT).

The problem of accuracy is still present even in monolingual SMT. A small model will not have high enough probability mass to be able to generalise well all the linguistic phenomena a good translation should encompass. The translation model will suffer from a low number of examples and thus might not be able to estimate the probabilities correctly. The unsupervised alignment model implemented in Moses using GIZA++ aligner (Och and Ney, 2003) will have rough statistics for the alignment estimation if computed from a small number of parallel sentences.

2.3 State-of-the-Art ATS System for Spanish

The current state-of-the-art text simplification system for Spanish (Saggion et al., 2015) was built under the Simplext project.[2] It employs a modular approach to TS, consisting of three main modules: a rule-based syntactic and lexical simplification modules (Drndarević et al., 2013); and a synonym-based lexical simplification module (Bott et al., 2012). According to the recent evaluation of the full Simplext system (Saggion et al., 2015), the system achieved human scores for grammaticality, meaning preservation, and simplicity comparable to those of the current state-of-the-art data-driven text simplification systems for English (Wubben et al., 2012; Angrosh and Siddharthan, 2014).

3 Methodology

The corpora, translation/simplification experiments, and the evaluation procedure are presented

[1] https://simple.wikipedia.org/wiki/Main_Page

[2] www.simplext.es

Version	Example
Original	Los expertos presentarán un informe de esta misión en la próxima reunión del Comité del Patrimonio Mundial, que tendrá lugar en Bahrein en junio de 2011.
Light	Los expertos presentarán un informe de*l estudio del estado de conservación de Pompeya* en la próxima reunión del Comité del Patrimonio Mundial, que *será* en Bahrein en junio de 2011.
Heavy	Los expertos presentarán un informe *sobre Pompeya* en la próxima reunión *sobre la cultura del mundo. Esta reunión será* en junio de 2011.

Table 1: Different levels of simplification (deviations from the original sentence are shown in italics)

in the next three subsections.

3.1 Corpora

In order to test the influence of the level of simplification in TS datasets ("heavy" or "light") on the system performance, we trained the standard PB-SMT (Koehn et al., 2003) and HIERO (Chiang, 2007) models in the Moses toolkit (Koehn et al., 2007) on two TS corpora:

1. *Heavy* – The TS corpus built under the Simplext project (Saggion et al., 2011), aimed at simplifying texts for people with intellectual disabilities. The original news stories were simplified manually by trained human editors, following detailed guidelines (Anula, 2007).

2. *Light* – The TS corpus consisting of various texts (some of which present in the *Heavy* corpus) and their manual simplifications obtained using only six main simplification rules (Mitkov and Štajner, 2014).

In both corpora, the sentence-alignment was manually checked and corrected where necessary. An example of an original sentence and its corresponding manual simplifications in the two corpora is given in Table 1.

3.2 SMT Models

In order to compare the impact of different SMT models (PB vs. HIERO) on the system performance, the language model (LM) and the test set (consisting of 47 sentence pairs from each of the two corpora) were kept the same for all systems. Ideally, the LM should be trained on a large corpus of "simplified" Spanish. However, as such a corpus has not been compiled yet, we trained the LM on a subset of the Europarl v7 Spanish corpus (Koehn, 2005) using the SRILM toolkit (Stolcke, 2002). In order to reduce the complexity of the sentences used for the training of the LM, we filtered out all sentences that contain more than 15 tokens. The sizes of the datasets used in the four experiments are given in Table 2.

Corpus	Model	Training	Dev.	Test
Light	PB-SMT	659	100	94
Light	HIERO	659	100	94
Heavy	PB-SMT	725	100	94
Heavy	HIERO	725	100	94

Table 2: SMT experiments

3.3 Evaluation

In order to obtain better insights into the potential problems in the SMT-based ATS, where the models are trained on the small datasets, we opted for human evaluation of the output in addition to the automatic evaluation (using the BLEU scores). Following the standard procedure for human evaluation of TS systems used in previous studies (Coster and Kauchak, 2011; Wubben et al., 2012; Drndarević et al., 2013), we asked human evaluators to assess, on a 1–5 scale (where the higher mark always denotes better output), three aspects of the presented sentences: grammaticality (G), meaning preservation (M), and simplicity (S).

We first asked thirteen annotators (8 native and 5 non-native with advanced knowledge of Spanish) to rate 20 original sentences and their corresponding simplifications (one manual and four automatic SMT-based) in order to directly compare the performances of the PB and HIERO models on both corpora. Next, we asked the annotators to rate another 20 original sentences and their corresponding simplifications (one manual and three automatic, out of which two were produced by the two best SMT systems and the third by the Simplext system) in order to directly compare the performances of the SMT systems with the state-of-the-art (rule-based) text simplification system for Spanish (Section 2.3).

We obtained a total of 260 human scores for each aspect-system-corpora combination in each of the two evaluation phases. The 40 original sentences for human evaluation (and their corre-

System	Corpus	S-BLEU	BLEU
PB	Light	**0.3742**	0.3374
	Heavy	0.3662	0.3313
HIERO	Light	**0.3718**	0.3336
	Heavy	0.2959	0.2718
Baseline		0.3645	0.3260

Table 3: Automatic evaluation

sponding simplified variants) were selected randomly from the test set under the criterion that they have been modified by at least two ATS systems. Every annotator was asked to rate all versions of the same original sentence (different versions of the same sentence were always shown in a random order). This allowed us to have a direct pairwise comparison of each pair of systems.

4 Results

The results of the automatic and human evaluations are presented in the next two subsections.

4.1 Automatic Evaluation

We compared the performances of the systems using two automatic MT evaluation metrics, the sentence-level BLEU score (S-BLEU)[3] and the document-level BLEU score (Papineni et al., 2002). As the baseline, we used the system which makes no changes to the input (i.e. output of the system is the original sentence). This seems as a natural baseline for this specific task (ATS), as all previous studies (Specia, 2010; Coster and Kauchak, 2011; Štajner, 2014) reported that their systems are overcautious, usually making only a few or no changes to the input sentence, and only slightly outperform this baseline. For calculating the S-BLEU and BLEU scores, we used the manual simplification ('gold standard') as the reference, and the original sentences and the outputs of the four systems as five corresponding hypotheses. The results are presented in Table 3.

The only two systems which significantly outperform the baseline in terms of the S-BLEU scores (0.05 level of significance; Wilcoxon signed rank test for repeated measures) are the systems trained and tuned on the *Light* corpus. The performance of the PB and HIERO systems trained and tuned on the *Heavy* corpus was not significantly different from the baseline.

[3]Sentence-level BLEU score (S-BLEU) differs from BLEU score only in the sense that S-BLEU will still positively score segments that do not have higher n-gram matching (n=4 in our setting) unless there is no unigram match; otherwise it is the same as BLEU.

Aspect		Heavy		Light		Manual
		PB	HIERO	PB	HIERO	
G	Mean	1.74	1.77	**4.03**	3.91	4.61
	Median	1	1	5	4	5
	Mode	1	1	5	5	5
M	Mean	1.98	1.93	**4.57**	4.40	3.62
	Median	1	1	5	5	4
	Mode	1	1	5	5	4
S	Mean	2.31	2.29	**2.99**	2.93	4.40
	Median	2	2	3	3	5
	Mode	1	1	2	2	5

Table 4: Phrase-based vs. hierarchical SMT

4.2 Human Evaluation

The results of the human evaluation of the PB and HIERO systems built using each of the two corpora (*Heavy* and *Light*) are given in Table 4. For each of the three aspects (G – grammaticality, M – meaning preservation, and S – simplicity), we present the mean, median and mode calculated on the 260 entries for each system-corpus combination.

As can be seen (Table 4), the systems built using the *Light* corpus were rated higher than those built using the *Heavy* corpus on all three aspects (especially pronounced for grammaticality and meaning preservation). The performances of the PB and HIERO models built using the *Heavy* corpus achieved almost the same scores, while the PB model was rated as slightly better than HIERO in the case when the models were built using the *Light* corpus (the differences in G and M scores were statistically significant at a 0.01 level of significance[4]). It is interesting to note that the meaning preservation (M) score was higher for both SMT-models built using the *Light* corpus than for the manual simplifications. This reflects the fact that manual simplification often relies on heavy paraphrasing and sometimes does not retain all information present in the original sentence (see the example in Table 9, Section 5).

Additionally, we calculated how many times: (1) the output of the systems built using the *Light* corpus was rated better than the output of the systems built using the *Heavy* corpus on the same test sentence (Table 5), and (2) the output of the HIERO models was rated better than the output of

[4]Statistical significance was measured in SPSS using the marginal homogeneity test which represent the extension of McNemar test from binary to multinominal response for two related samples.

Comparison	HIERO	PB
G(Light) > G(Heavy)	221	228
G(Light) = G(Heavy)	36	29
G(Light) < G(Heavy)	3	3
M(Light) > M(Heavy)	225	230
M(Light) = M(Heavy)	31	28
M(Light) < M(Heavy)	4	2
S(Light) > S(Heavy)	119	119
S(Light) = S(Heavy)	88	84
S(Light) < S(Heavy)	53	57

Table 5: Impact of the corpora used

Comparison	Light	Heavy
G(HIERO) > G(PB)	12	39
G(HIERO) = G(PB)	216	182
G(HIERO) < G(PB)	32	39
M(HIERO) > M(PB)	7	36
M(HIERO) = M(PB)	217	179
M(HIERO) < M(PB)	36	45
S(HIERO) > S(PB)	22	40
S(HIERO) = S(PB)	219	171
S(HIERO) < S(PB)	19	49

Table 6: Impact of the model used

the PB models on the same test sentence (Table 6). The results of these comparisons confirmed that both models (HIERO and PB) achieve better performances if they are built using the *Light* corpus instead of using the *Heavy* corpus (Table 5). It also seems that the PB model generates more grammatical sentences and better preserves the original meaning than the HIERO model when trained the *Light* corpus, while both models lead to similar performances when trained on the *Heavy* corpus (Table 6).

5 Comparison with the State of the Art

The results of the human evaluation of 20 original sentences and their four corresponding simplified versions (Table 7) indicate that the output of the SMT-based systems is more grammatical and preserves the meaning better than the output of Simplext, at the cost of being less simple.[5] The pairwise comparison of 260 sentences (Table 8) confirmed those findings.

Aspect		PB	HIERO	Simplext	Manual
G	Mean	3.68	**3.86**	3.49	4.47
	Median	4	4	4	5
	Mode	4	4	4	5
M	Mean	4.17	**4.37**	3.95	3.17
	Median	4	5	4	3
	Mode	5	5	5	4
S	Mean	2.60	2.61	**2.80**	4.42
	Median	3	3	3	5
	Mode	3	2	3	5

Table 7: Comparison with the state of the art

Comparison	PB	HIERO
G(SMT-based) > G(Simplext)	96	104
G(SMT-based) = G(Simplext)	90	99
G(SMT-based) < G(Simplext)	76	57
M(SMT-based) > M(Simplext)	92	103
M(SMT-based) = M(Simplext)	109	113
M(SMT-based) < M(Simplext)	59	44
S(SMT-based) > S(Simplext)	59	55
S(SMT-based) = S(Simplext)	96	105
S(SMT-based) < S(Simplext)	95	100

Table 8: Comparison with Simplext

An example of an original sentence, its manual simplification ("gold standard"), and its automatic simplifications by three different systems (PB, HIERO, and Simplext) are given in Table 9. In this example, both SMT-based systems perform two lexical simplifications: (1) "galardón (*award*) is replaced with "premio" (*prize*), and (2) "concede" (*concede*) is replaced with "da" (*gives*). These lexical substitutions lead to a sentence which is simpler than the original and preserves the original meaning. In the same example, the Simplext system performs two syntactic simplifications by splitting the original sentence into three new sentences, out of which only one (the second) is grammatical and preserves the original meaning. The first of the three new sentences is grammatical but changes the original meaning, while the third one is neither grammatical, nor preserves the original meaning. Additionally, in this example, the Simplext system does not lexically simplify the original sentence. The manually simplified sentence is, as expected, the simplest and most grammatical. However, it represents a very strong paraphrase of the original sentence which does not preserve the original meaning faithfully and is, therefore, penalised with the lowest score for the meaning preservation out of all four simplification variants.

6 Error Analysis

In order to better understand the shortcomings of the SMT-based systems (and the phrase-based approach to ATS using small size corpora, in general), we performed manual error analysis of all sentences for which the SMT-based systems received lower scores than the Simplext system (on

[5] All differences, except the S score for the PB and HIERO models, are statistically significant at a 0.05 level of significance.

Version	Example	G	M	S
Original	Este galardón, dotado con 20.000 euros, lo concede el Ministerio de Cultura para distinguir una obra de autor español escrita en cualquiera de las lenguas oficiales y editada en España durante 2009.	4.77	5.00	2.84
PB/HIERO	Este *premio*, dotado con 20.000 euros, lo *da* el ministerio de cultura para distinguir una obra de autor español escrita en cualquiera de las lenguas oficiales y editada en España durante 2009.	4.15	4.77	3.15
Simplext	Este galardón lo concede el Ministerio de Cultura para distinguir una obra de autor español *durante el año 2009. El galardón está* dotado con 20.000 euros. *El autor está* escrita en cualquiera de las lenguas oficiales y editada en España.	3.54	3.77	2.92
Manual	Este premio es para un autor español que escriba en español, catalán, vasco o gallego.	4.85	3.31	4.85

Table 9: An example of an original sentence, its manual simplification, and its automatic simplifications generated by three different ATS systems (deviations from the original sentences are shown in italics; columns 'G', 'M', and 'S' contain mean value of the scores for grammaticality, meaning preservation, and simplicity obtained from all thirteen annotators)

average). In all of those cases when the SMT-based systems scored lower than the Simplext system the reason was one (or both) of the following: the system performed one wrong lexical substitution which led to a low grammaticality score; and/or the system did not perform sentence splitting and the Simplext system did. Table 10 contains three such examples.

In the first example (1), the SMT-based systems applied an incorrect lexical substitution, replacing the word "informó" (*informed*) with "gracias" (*thanks*). That led to an ungrammatical output of the system and the lower total score. The same word was correctly simplified by the Simplext system using the word "dijo" (*said*) instead. The Simplext system additionally performed a sentence splitting. During that process, the name of the university at which the writer graduated has been replaced with the name of the writer, which changed the original meaning of the sentence. However, this did not lead to an ungrammatical output (as opposed to the wrong lexical substitution performed by the SMT-based models), and the Simplext system thus obtained better scores for grammaticality (G) and simplicity (S), and a lower score for meaning preservation (M) than the SMT-based systems.

In the second example (2), the SMT-based systems performed one good lexical simplification (which was not performed by the Simplext system) by replacing the word "aseguró" (*assured*) with the word "dice" (*says*). However, our systems also applied one incorrect lexical simplification which, although it did not change the original meaning of the sentence, led to the ungrammatical output. In the same example, the Simplext system correctly split the original sentence into two shorter sentences and performed one correct lexical simplification. The changes made by the Simplext system led to a small grammatical issue ("poner le" should be written together), but this did not significantly influence grammaticality score (G).

The third example (3) illustrates the case in which the Simplext system was rated better than the SMT-based systems because it performed a sentence splitting when the SMT-based systems did not. At the same time, the SMT-based systems applied one correct lexical simplification. The same word was left unchanged by the Simplext system. However, it appears that human evaluators tend to give a higher simplicity score to the system which performs sentence splitting than to the system which performs lexical simplification (in the case that each of the systems performs only one of the two possible modifications).

7 Conclusions and Future Work

In this paper, we presented the results of the phrase-based (PB) and hierarchical (HIERO) SMT models for ATS, built using two small TS corpora. One corpus contained "heavy" simplifications, and the other "light" simplifications. The direct comparison of the systems' performances, based on an extensive human evaluation, indicated that both models (PB and HIERO) achieve similar performances if they are built using the same corpus (either *Heavy* or *Light*). The results of the human evaluation also showed that SMT-based models built using the *Light* corpus generate sentences that are more grammatical and preserve the meaning better, but are less simple than those generated

Version	Example	G	M	S
(1) Original	Castellet (Barcelona, 1926), escritor, crítico literario y editor, estudió en la Universidad de Barcelona, donde se graduó en Derecho, según informó el Ministerio de Cultura.	4.85	5.00	3.46
(1) HIERO, PB	Castellet (Barcelona, 1926), escritor, crítico literario y editor, estudió en la Universidad de Barcelona, donde se graduó en Derecho, según *gracias* el Ministerio de Cultura.	3	3.54	3.15
(1) Simplext	Castellet, escritor, crítico literario y editor, estudió en la Universidad de Barcelona. *En Castellet se licenció* en Derecho, según *dijo* el Ministerio de Cultura.	4.69	3.46	3.53
(2) Original	El presidente del Grupo Planeta, José Manuel Lara, aseguró en el Foro de la Nueva Cultura que el problema de la piratería en España es"grave" y"preocupante" y la sociedad "debe tomar conciencia para ponerle coto".	4.46	5.00	2.77
(2) HIERO, PB	El presidente del Grupo Planeta, José Manuel Lara, *dice* en el Foro de la Nueva Cultura que el problema de la piratería en España es "grave" y "preocupante" y la sociedad "*hay* tomar conciencia para ponerle coto".	3.23	4.31	2.46
(2) Simplext	El presidente del Grupo Planeta, José Manuel Lara, aseguró en el Foro de la Nueva Cultura que el problema de la piratería en España es "grave" y "preocupante". *L*a sociedad "debe tomar conciencia para poner *le límite*".	4.23	4.77	3.38
(3) Original	Sin embargo, el terrorismo, que aparece en cuarto lugar (19%), registra la cota más baja de toda la serie desde 2004, experimentando una caída de 12 puntos respecto del Sociómetro de mayo.	4.38	5.00	3.15
(3) HIERO, PB	Sin *pero*, el terrorismo, que *sale* en cuarto lugar (19%), registra la cota más baja de toda la serie desde 2004, experimentando una caída de 12 puntos respecto del Sociómetro de mayo.	3.08	3.92	2.23
(3) Simplext	Sin embargo, el terrorismo*,* registra la cota más baja de toda la serie desde *el año* 2004, experimentando una caída de 12 puntos respecto del Sociómetro de mayo. *Este terrorismo* aparece en cuarto lugar.	3.08	3.77	2.92

Table 10: Three examples of the original sentences and their automatic simplifications generated by our systems and the Simplext system (deviations from the original sentences are shown in italics; the columns 'G', 'M', and 'S' contain mean value of the scores for grammaticality, meaning preservation, and simplicity obtained from all thirteen annotators)

by the Simplext system.

We acknowledge that the fact that we built the language models using the Europarl corpus which is not a good representative of "simplified" language (despite our efforts to filter out complex sentences) is probably one of the main reasons why the SMT-based systems are not able to generate sentences as simple as those generated by Simplext. Our future work will thus focus on finding better strategies for filtering out complex sentences from the Europarl corpus (e.g. using just those sentences with certain simple syntactic structures, and those with simple and frequently used words).

Acknowledgements

The research described in this paper was partially funded by the project SKATER-UPF-TALN (TIN2012-38584-C06-03), Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain, the project ABLE-TO-INCLUDE (CIP-ICT-PSP-2013-7/621055), and by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University. We are also grateful to all the annotators for their evaluation efforts, and the anonymous reviewers for their constructive comments.

References

Sandra Maria Aluísio and Caroline Gasperin. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (YIWCALA)*, pages 46–53. ACL.

M.A. Angrosh and A. Siddharthan. 2014. Text simplification using synchronous dependency grammars: Generalising automatically harvested rules. In *Proceedings of the 8th International Natural Language Generation Conference (INGL)*, pages 16–25.

Alberto Anula. 2007. Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2013. Transforming Com-

plex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural*, 50:61–68.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING*, pages 357–374.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for french. In *Proceedings of the EACL Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden*, pages 47–56.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

William Coster and David Kauchak. 2011. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9. ACL.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 488–500.

Ann Irvine and Chris Callison-Burch. 2013. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, volume 1, pages 48–54. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. ACL.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit*.

Ruslan Mitkov and Sanja Štajner. 2014. The Fewer, the Better? A Contrastive Study about Ways to Simplify. In *Proceedings of the COLING workshop on Automatic Text Simplification – Methods and Applications in the Multilingual Society (ATSMA), Dublin, Ireland*, pages 30–40.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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 of the Association for Computational Linguistics (ACL)*, pages 311–318.

Horacio Saggion, Elena Gómez Martínez, Alberto Anula, Lorena Bourg, and Esteban Etayo. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 30–39. Springer-Verlag Berlin, Heidelberg.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Sanja Štajner. 2014. Translating sentences from 'original' to 'simplified' spanish. *Procesamiento del Lenguaje Natural*, 53:61–68.

Kristian Woodsend and Mirella Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 409–420.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1, pages 1015–1024.

Zhemin Zhu, Delphine Berndard, and Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1353–1361.

Appendix A: Scoring Instructions Given to the Annotators

	Grammaticality (G)
5	Grammatically correct sentence
4	One or two typos (not capitalised first letter of the sentence, ''s' separated from the noun, missing comma, etc.)
3	One incorrect construction but the sentence still has a meaning (missing preposition in a phrasal verb, transitive instead of intransitive verb or vice versa, use of animate instead of inanimate object or vice versa, etc.)
2	A few incorrect constructions of the above type, or a combination of a typo and an incorrect construction, but the sentence is still meaningful
1	So many mistakes (or such a mistake) that the sentence is grammatically incorrect and completely meaningless
	Meaning preservation (M)
5	The two sentences have exactly the same meaning
4	The meanings of the two sentences differ just in a nuance or some minimal addition of a world knowledge
3	The two sentences do not mean exactly the same, but the main point is the same
2	The meanings of the two sentences differ, but they are not opposite
1	The meanings of the two sentences are opposite
	Simplicity (S)
5	Very simple (all words are short, frequent, and used with their most commonly used meaning)
4	Simple (a few longer words, but still frequent and used with their most commonly used meaning)
3	A few difficult words or phrases, but the overall meaning of the sentence is clear
2	Quite a few difficult words or phrases which makes it difficult to understand the main meaning of the sentence
1	Very difficult to understand (many difficult words and phrases, not used with their most commonly used meaning)