Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 135-142 recibido 25-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Adapting Text Simplification Decisions to Different Text Genres and Target Users

Adaptación de algoritmos de toma de decisiones de simplificación de textos a diferentes corpus y audiencias

Sanja Štajner
University of Wolverhampton, UK
sanjastajner@wlv.ac.uk

Horacio Saggion
Universitat Pompeu Fabra, Spain
horacio.saggion@upf.edu

Resumen: Hemos analizado las alineaciones a nivel de oración de dos corpus paralelos de textos originales y sus simplificaciones creados con diferentes objetivos. Hemos clasificado las alineaciones que se observan y diseñado un algoritmo de clasificación capaz de predecir si las oraciones de un texto serán eliminadas, segmentadas, o transformadas durante el proceso de simplificación. Hemos realizado una evaluación cruzada en cada uno de los corpus así como una evaluación en la cual se entrena en algoritmo en un corpus y se lo evalúa en el otro.
Palabras clave: Simplificación de textos, clasificación de oraciones, adaptación de métodos

Abstract: We investigate sentence deletion and split decisions in Spanish text simplification for two different corpora aimed at different groups of users. We analyse sentence transformations in two parallel corpora of original and manually simplified texts for two different types of users and then conduct two classification experiments: classifying between those sentences to be *deleted* and those to be *kept*; and classifying between sentences to be *split* and those to be left *unsplit*. Both experiments were first run on each of the two corpora separately and then run by using one corpus for the training and the other for testing. The results indicated that both sentence decision systems could be successfully trained on one corpus and then used for a different text genre in a text simplification system aimed at a different target population.
Keywords: Text simplification, sentence classification, method adaptation

1 Introduction

Since the late nineties several initiatives which proposed guidelines for producing plain, easy-to-read and more accessible documents have emerged, e.g. "The Plain Language Action and Information Network (PLAIN)"[1], "Make it Simple, European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability" (Freyhoff et al., 1998), "Am I making myself clear? Mencap's guidelines for accessible writing"[2], and "Web content accessibility guidelines"[3]. All these initiatives increased the interest in the use of natural language processing in the development of assistive technologies and automatic text simplification, as it is clear that manual simplification cannot match the rate of production of texts, particularly of newswire texts which are being constantly generated.

The first systems aimed at automatic text simplification were rule-based, e.g. (Chandrasekar, 1994; Devlin, 1999; Devlin and Unthank, 2006). Syntactic simplification modules usually consisted of a set of rules which are recursively applied to each sentence as long as it is possible. Lexical simplification modules were traditionally based on substitution of difficult infrequent words with their simpler synonyms.

With the emergence of Simple English Wikipedia[4] the approaches to automatic text simplification became more data-driven. Biran et al. (2011) and Yatskar et al. (2010), apply an unsupervised method for learning pairs of complex and simple synonyms from a corpus of texts from the original Wikipedia

[1] http://www.plainlanguage.gov/

[2] http://november5th.net/resources/Mencap/Making-Myself-Clear.pdf

[3] http://www.w3.org/TR/WCAG20/

[4] http://simple.wikipedia.org

ISSN 1135-5948

and Simple English Wikipedia. Coster and Kauchak (2011a; 2011b) address the problem of text simplification as an English-to-English translation problem. They use the standard machine translation tools trained on the parallel corpus of aligned sentences from original and Simple English Wikipedia, to build an automatic text simplification system. Although the results show that the machine translation approach to text simplification works well for English, the same approach cannot be applied to other languages, as Simple Wikipedia does not exist for many languages (Spanish among them). Another limitation is that, although it imposes the use of Basic English vocabulary, shorter sentences and simpler grammar, Simple English Wikipedia does not follow easy-to-read guidelines for writing for people with cognitive disabilities. Therefore, it may not represent a good training material for text simplification for this target audience.

The compilation of a parallel corpus of original and manually simplified texts for specific target audiences (e.g. people with learning or language disabilities) is both time-consuming and expensive (involving special training for human annotators and adaptation of easy-to-read guidelines for a specific language and target population). Therefore, it would be important to investigate whether the simplification systems (or some of their components) developed for one specific target population and text genre could also be used for text simplification aimed at other target populations and different text types – a problem never addressed before. This paper fills that gap, exploring whether sentence deletion and split decisions learned from a parallel corpus of news texts compiled for the needs of a specific user group could be used for different user groups and text genres. As shown in this paper, the decisions learned can be transferred to a new corpus if an appropriate learning algorithm is used.

The reminder of the paper is organised as follows: Section 2 presents the most relevant previous work on the topic of sentence decisions in text simplification; Section 3 describes the corpora used in this study and presents the results of the initial analysis of detected sentence transformations in both corpora; Section 4 introduces the features and the settings for the two classification experiments; Section 5 presents and discusses the results of the classification experiments; and Section 6 draws attention to the main findings of the presented study and offers possible directions for future work.

2 Related Work

Various studies have described necessary transformations to be included in an automatic text simplification system for the English language. They analysed the parallel corpora of original and manually simplified texts aimed at different target audiences: (1) for children (Bautista et al., 2011), using Encyclopedia Britannica and Britannica Elemental (Barzilay and Elhadad, 2003); (2) for language learners (Petersen and Ostendorf, 2007), using original and abridged texts from Literacyworks[5]; and (3) for audiences with various reading difficulties (Biran, Brody, and Elhadad, 2011; Yatskar et al., 2010; Coster and Kauchak, 2011a; Coster and Kauchak, 2011b), using original and Simple English Wikipedia.

Petersen and Ostendorf (2007) reported that 30% of sentences were completely eliminated, while 19% of sentences were split into two or more sentences by the human editors while simplifying texts for language learners in English. Caseli et al. (2009) showed sentence splitting to be the second most frequent simplification operation, present in 34% of the original sentences (straight after lexical substitution present in 46% of the sentences), while only 0.28% of sentences were completely eliminated, during the manual simplification of text for people with low literacy levels in Brazilian Portuguese. Štajner et al. (2013) performed a similar analysis on a small corpus of original and manually simplified texts (37 text pairs) in Spanish, aimed at people with cognitive disabilities. They reported sentence deletion and sentence splitting as being almost equally present simplification operations (21% and 23% of original sentences, respectively).

Motivated by those previous studies, this article: (1) analyses the types of applied manual transformations in Spanish text simplification aimed at two different target populations: people with intellectual disabilities (Down's syndrom), and people with autism spectrum disorders (ASD); and (2) proposes

[5]http://literacynet.org/cnnsf/index_cnnsf.html

the algorithms for classification of original sentences into those which should be *deleted*, *split*, and left *largely unchanged*.

More importantly, this study goes one step further by testing whether the sentence classification system built on one specific text genre and aimed at one specific target population can successfully be applied in other text genres and for different target populations for which parallel corpora of original and manually simplified texts may not exist. To the best of our knowledge, this is the first study addressing the problem of method adaptation in text simplification.

3 Corpora

The main corpus (Corpus A henceforth) used in the experiments contains 195 original and manually simplified news articles in Spanish (a total of 1118 orignal sentences), provided by the Spanish news agency Servimedia[6] and compiled under the Simplext project[7] (Saggion et al., 2011). Simplifications have been applied by trained human editors, familiar with the particular needs of a person with cognitive disabilities (Down's syndrom) and following a series of easy-to-read guidelines. The corresponding pairs of original and simplified texts were first sentence aligned using an automatic alignment tool (Bott and Saggion, 2011) and then manually post-edited in order to provide 100% accurate sentence alignment.

The second corpus (Corpus B henceforth) is significantly smaller and comprises 25 original and manually simplified texts (a total of 320 original sentences) of different genres: literature, news, health, general culture and instructions. It was compiled under the FIRST project[8] (Orasan, Evans, and Dornescu, 2013). Texts were manually simplified by five experts who have experience of working with people with autism, having in mind the particular needs of this target population. The corresponding pairs of original and simplified texts were sentence aligned manually, thus ensuring alignment to be 100% accurate.

3.1 Sentence Transformations

By automatically processing the aligned sentences in Corpus A it was found that: (1) the original sentence was neither split nor deleted ("1-1" alignment) in 566 cases; (2) the original sentence was split into two or more sentences ("1-n" alignment) in 358 cases; and (3) the original sentence was completely deleted ("1-0" alignment) in 186 cases. The same analysis of the aligned sentences in Corpus B (total of 305 sentences) revealed that: (1) the original sentence was neither split nor deleted ("1-1" alignment) in 192 cases; (2) the original sentence was split into two or more sentences ("1-n" alignment) in 70 cases; and (3) the original sentence was completely deleted ("1-0" alignment) in 43 cases (Table 3.1).

Transformation		Corpus A	Corpus B
"1-0"	deleted	186 (17%)	43 (14%)
"1-n"	split	358 (32%)	70 (23%)
"1-1"	same	275 (25%)	178 (58%)
	reduced	291 (26%)	14 (5%)
Total ("1-x")		1110 (100%)	305 (100%)

Table 1: Corpus analysis

More detailed analysis of "1-1" aligned sentences, revealed that in many cases original sentences were significantly longer than their simplified versions, thus indicating that certain parts of the original sentences were omitted during the simplification process, as in the following example of original (1) and its corresponding simplified sentence (2):

1. *"El Premio de la Cinematografía y de las Artes Audiovisuales está destinado a recompensar la aportación más sobresaliente en el ámbito cinematográfico español puesta de manifiesto a través de una obra hecha pública durante 2009, o de una labor profesional desarrollada durante ese mismo año."*

2. *"El Premio Nacional de Cine se da a la mejor película o trabajo del año 2009."*

Therefore, the "1-1" aligned sentences were further divided into two groups: *same* – those sentences which were only slightly modified (the difference between number of words in the original and simplified sentence is less than ten words); and *reduced* – those sentences whose lengths were significantly reduced during the simplification (the difference between number of words in the original and simplified sentence is ten or more words). Unlike Corpus A, which contains a

[6] http://www.servimedia.es/
[7] http://www.simplext.es/
[8] http://first-asd.eu/

significant number of *reduced* sentences, Corpus B contains only 14 cases of these sentences (Table 3.1). These sentences were thus excluded from Corpus B in all classification experiments.

Analysis of sentence transformations in both corpora revealed an additional, frequently occurring type of transformation – *enlarged* sentences (simplified sentence is at least ten words longer than its original). All of those were the result of adding a definition of a complex term, as in the following example of original (1) and its corresponding simplified sentence (2):

1. *"He visitado cientos de mundos, he sido dama victoriana, rey medieval y bucanero."*
2. *"Al leer novelas he visitado cientos de mundos, he sido una dama de la época victoriana (época transcurrida entre 1837 y 1901), un rey medieval (de la época transcurrida entre el siglo V y el siglo XV) y un bucanero (un pirata que en los siglos XVII y XVIII robaba las posesiones españolas de ultramar)."*

These *enlarged* sentences did not significantly differ from the *same* sentences in terms of the features used in this paper. Therefore, they were counted as occurences of the *same* sentences and treated as such in all classification experiments.

3.2 Additional Types of Sentence Transformations

While the aforementioned sentence transformations were expected to be found in the corpora, it was surprising to discover that in several cases (four in Corpus A and six in Corpus B) two original sentences were merged into one simplified sentence ("2-1" alignment), as in the following pair of two original sentences (1) and their corresponding simplified sentence (2):

1. *"El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el 65% de los atentados a profesionales sanitarios. Y el grupo de edad más castigado, el que va desde los 46 a los 55 años."*
2. *"Los médicos que sufren más ataques son los de alrededor de 50 años y los que trabajan en centros médicos pequeños."*

In addition to the very frequent type of *enlarged* sentences, in several cases, even whole sentences were added as a definition. Especially interesting are the cases in which the addition of a definition (in a separate sentence) occurred simultaneously with sentence splitting as in the following case of original sentence (1) and its corresponding simplified paragraph (2) in Corpus B:

1. *"Este nombre se da a una mezcla gaseosa, líquida y sólida de hidrocarburos, que se ha encontrado en depósitos de rocas sedimentarias, en diferentes proporciones y en distintos lugares de la Tierra."*
2. *"El petróleo es una mezcla: Gaseosa, líquida y sólida de hidrocarburos. Los hidrocarburos son una mezcla de hidrógeno y carbono. El petróleo se ha encontrado en depósitos de rocas sedimentarias (en capas de rocas), en diferentes cantidades y en diferentes lugares de la Tierra."*

These *merged* and *added* sentences were not used in any of the classification experiments presented in this paper.

4 Experimental Settings

The corpora were parsed with state-of-the-art Connexor's Machinese parser[9] and the features (Table 3.2) were automatically extracted using the parser's output. Each sentence is represented as vector of 24 features inspired by the works of Štajner et al. (2013), Gasperin et al. (2009), Petersen and Ostendorf (2007), and Drndarevic and Saggion (2012). Features 1-19 and 21-22 count the number of occurrences of the feature in the sentence (e.g. feature 1 counts how many verbs the sentence has while feature 10 counts the number of determiners in the sentence). Feature 20 represents the position of the sentence in the text.

All classification experiments were conducted in Weka Experimenter (Witten and Frank, 2005), employing four different classification algorithms: Naive Bayes (John and Langley, 1995); SMO (Weka implementation of Support Vector Machines) with normalisation and using poly kernels (Keerthi et al.,

[9] www.connexor.eu

#	Code	Feature	#	Code	Feature	#	Code	Feature
1	*v*	verb	9	*pron*	pronoun	17	*main*	head of the verb phrase
2	*ind*	indicative	10	*det*	determiner	18	*nh*	head of the noun phrase
3	*sub*	subjunctive	11	*n*	noun	19	*advl*	head of the adverbial phrase
4	*inf*	infinitive	12	*prep*	preposition	20	*sent*	position of the sentence
5	*pcp*	participle	13	*cc*	coord. conj.	21	*punc*	punctuation marks
6	*ger*	gerund	14	*cs*	subord. conj.	22	*num*	numerical expressions
7	*adj*	adjective	15	*prem*	pre-modifier	23	*char*	sentence length in characters
8	*adv*	adverb	16	*postm*	post-modifier	24	*words*	sentence length in words

Table 2: Feature set

Classifier	Corpus A			Corpus B			A tested on B			B tested on A		
	P	R	F	P	R	F	P	R	F	P	R	F
SMO*	0.69	0.83	0.76	0.76	0.87	0.81	0.76	0.87	0.81	0.69	0.83	0.76
NB	0.76	0.81	0.78	0.82	0.62	0.68	0.80	0.83	0.81	0.71	0.67	0.69
JRip	**0.79**	**0.83**	**0.80**	0.81	0.85	0.82	0.76	0.75	0.75	**0.86**	**0.84**	**0.76**
J48	0.77	0.79	0.77	**0.84**	**0.87**	**0.84**	0.76	0.70	0.73	**0.79**	**0.83**	**0.76**

Table 3: Results of the classification between deleted and kept sentences (Key: Corpus A = 10-fold cross-validation with ten repetitions using only corpus A; Corpus B = 10-fold cross-validation with ten repetitions using only corpus B; A on B = training set: corpus A, test set: corpus B; B on A = training set: corpus B, test set: corpus A)

2001; Platt, 1998), JRip (Cohen, 1995), and J48 (Weka implementation of C4.5) (Quinlan, 1993). The experiments were the following:

- Experiment I: Classification between *deleted* ("1-0") and *kept* ("1-1" and "1-n") sentences;
- Experiment II: Classification between *split* and *unsplit* (*same*) sentences.

5 Results and Discussion

Results for each of the experiments are presented and discussed separately in the next two subsections (Sections 5.1 and 5.2).

5.1 Sentence Deletion

The weighted average P (precision), R (recall), and F (F-measure) for each classifier and each setup are given in Table 3.2. It is important to note that the P, R, and F values for the class *deleted* in SMO were 0, and thus can be taken as a baseline which does not delete any sentences (majority class). For each experiment, the results of the classifier which outperformed the baseline (row 'SMO*' in Table 3.2) on all three measures (P, R, and F) are shown in bold.

JRip achieved a significantly better precision (P) than SMO in the cross-validation setup on Corpus A, and when trained on Corpus B and tested on Corpus A. However, when trained on Corpus A and tested on Corpus B, the JRip classifier had a significantly lower performance (P, R, and F) than when used with a 10-fold cross-validation setup only on Corpus A. In general, the 10-fold cross-validation setup on each of the corpora separately, achieved better classification results than the setup with training on one corpus and testing on the other. None of the three classifiers (NB, JRip, and J48) outperformed the baseline (SMO) on any of the two setups ('A on B' and 'B on A') in terms of F-measure, although JRip and J48 achieved a significantly better precision (P) than the baseline.

Two additional experiments were conducted in order to explore whether: (1) elimination of the *reduced* sentences from the Corpus A; or (2) reduction of the feature set to the subset of best features (obtained by using the CfsSubsetEval attribute selection algorithm in Weka (Hall and Smith, 1998)), could improve the classification accuracy. Given that the results of these experiments were not significantly different from the results of the initial experiments (Table 3.2), they are not presented here.

Previous works on deletion decisions in

Classifier	Corpus A			Corpus B			A tested on B			B tested on A		
	P	R	F	P	R	F	P	R	F	P	R	F
SMO	**0.94**	**0.93**	**0.93**	**0.94**	**0.93**	**0.93**	0.94	0.94	0.94	0.94	0.94	0.94
NB	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.93	0.93	0.94	0.93	0.93
JRip	0.91	0.90	0.91	0.91	0.90	0.91	0.94	0.94	0.94	0.94	0.94	0.94
J48	0.91	0.91	0.91	0.91	0.91	0.91	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**

Table 4: Results of the classification between split and unsplit sentences (Key: Corpus A = 10-fold cross-validation with ten repetitions using only corpus A; Corpus B = 10-fold cross-validation with ten repetitions using only corpus B; A on B = training set: corpus A, test set: corpus B; B on A = training set: corpus B, test set: corpus A)

Spanish using cross-validation achieved F-scores of 0.79 (Drndarević and Saggion, 2012), and 0.82 (Štajner, Drndarević, and Saggion, 2013). We therefore consider the performance of our classification algorithms and feature set reasonable, in spite of not being directly comparable to those previous works because of differences in corpus characteristics.

5.2 Sentence Splitting

For the experiment on classification between *split* and *unsplit* sentences, the *reduced* and *deleted* sentences were excluded from both corpora. The decision not to include *reduced* sentences into either of the two classes (*split* and *unsplit*) arose from the nature of the *reduced* sentences. They could be interpreted as sentences which were first split and then one part was deleted and the second maintained. Therefore, it is expected that the *reduced* sentences contain markers of all three other types of sentences – *deleted*, *split*, and *same*. Also, the percentage of *reduced* sentences in each of the corpora was very unbalanced (Table 3.1 in Section 3).

The results of this classification experiment (Table 4) were quite surprising. All classification algorithms achieved better performances when trained on one corpus and tested on the other corpus. This was particularly accentuated in the case of the J48 classification algorithm which achieved the F-measure of 0.96 in both setups – 'A on B' and 'B on A'. The Support Vector Machines (SMO) performed as the best classifier on each of the corpora separately (columns 'Corpus A' and 'Corpus B' in Table 4). Naive Bayes achieved very similar results as the SMO classifier in all setups. The J48 classifier (Weka implementation of C4.5 decision tree classifier) significantly outperformed all three other classifiers in 'A on B' and 'B on A' setups. Note that a baseline that choses the majority case (split for corpus A and non-split for corpus B) would have obtained F=0.56 on corpus A, F=0.43 on corpus B. Previous work on split decisions by Gasperin et al. (2009), although not directly comparable to ours because of the different language and corpus, achieved an F-score of 0.80. Štajner et al. (2013) achieved an F-measure of 0.92 for the same task on a smaller portion of Corpus A, using a slightly different set of features. We therefore consider the performance of our classifier and set of features on our datasets acceptable.

6 Conclusions and Future Work

In this paper we addressed the issue of sentence deletion and split decisions as a first step in building an automatic text simplification system for Spanish. More particularly, we investigated the adaptability of these decisions across different text genres and two different target populations.

The initial analysis of sentence transformations in two corpora containing different text genres and aimed at different target users revealed some interesting differences in simplification strategies which were applied by human annotators in these two cases. Furthermore, it revealed different distribution of those sentence transformations which were present in both corpora.

The classification of original sentences into those to be *deleted* and those to be *kept* achieved better accuracy when performed on each of the corpora separately using 10-fold cross-validation setup than when trained on one corpus and tested on the other. It also indicated the JRip and J48 classifiers as being the most suitable for this task (out of the four classifiers applied).

The classification of original sentences into those to be *split* and those to be left *unsplit* led to surprising results. All four classifiers achieved better accuracies when trained on one corpus and tested on the other than when performed on each of the corpora separately in a 10-fold cross-validation setup. The difference in the classifier performance between the two setups was most pronounced in the case of the J48 (decision tree) classifier.

In the future, we plan to perform similar experiments on a larger number of corpora aimed at other target populations – second language learners, children, and users with different reading and learning disabilities. The main goal would be to discover how much of the methodology and system components could be shared between the automatic text simplification systems aimed at different target users (and different text genres).

Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development (FIRST 287607). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. This work is supported by an Advanced Research Fellowship from Programa Ramón y Cajal (RYC-2009-04291) and by the project SKATER: Scenario Knowledge Acquisition – Knowledge-based Concise Summarization (TIN2012-38584-C06-03) , Ministerio de Economía y Competitividad, Secretaria de Estado de Investigaciṕn, Desarrollo e Innovación, Spain.

References

Barzilay, R. and N. Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bautista, S., C. León, R. Hervás, and P. Gervás. 2011. Empirical identification of text simplification strategies for reading-impaired people. In *European Conference for the Advancement of Assistive Technology*.

Biran, O., S. Brody, and N. Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Bott, Stefan and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Caseli, H. M., T. F. Pereira, L. Specia, T. A. S. Pardo, C. Gasperin, and S. M. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *10th Conference on Intelligent Text PRocessing and Computational Linguistics (CICLing 2009)*.

Chandrasekar, R. 1994. *A Hybrid Approach to Machine Translation using Man Machine Communication*. Ph.D. thesis, Tata Institute of Fundamental Research/University of Bombay, Bombay.

Cohen, W. 1995. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

Coster, W. and D. Kauchak. 2011a. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1–9.

Coster, W. and D. 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. Portland, Oregon, USA: Association for Computational Linguistics*, pages 665–669.

Devlin, S. 1999. *Simplifying natural language text for aphasic readers*. Ph.D. thesis, University of Sunderland, UK.

Devlin, S. and G. Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th international*

ACM SIGACCESS conference on Computers and accessibility, Assets '06, pages 225–226, New York, NY, USA. ACM.

Drndarević, B and H. Saggion. 2012. Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *SEPLN Journal*, 49.

Freyhoff, G., G. Hess, L. Kerr, B. Tronbacke, and K. Van Der Veken, 1998. *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Gasperin, C., L. Specia, T. Pereira, and S.M. Aluisio. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA-2009), Bento Gonçalves, Brazil.*, pages 809–818.

Hall, M. A. and L. A. Smith. 1998. Practical feature subset selection for machine learning. In C. McDonald, editor, *Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98*, pages 181–191. Berlin: Springer.

John, G. H. and P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

Orasan, C., R. Evans, and I. Dornescu. 2013. Text Simplification for People with Autistic Spectrum Disorders. In D. Tufis, V. Rus, and C. Forascu, editors, *Towards Multilingual Europe 2020: A Romanian Perspective*. Romanian Academy Publishing House, Bucharest, pages 187–312.

Petersen, S. E. and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*.

Platt, J. C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*.

Quinlan, R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Saggion, H., E. Gómez-Martínez, A. Anula, L. Bourg, and E. Etayo. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del Lenguaje Natural*, 46.

Štajner, S., B. Drndarević, and H. Saggion. 2013. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computacion y Systemas*, 17(2):251–262.

Witten, I. H. and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Yatskar, M., B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.