Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification

Sanja Štajner[1], Biljana Drndarević[2], and Horacio Saggion[2]

[1]Research Group in Computational Linguistics, University of Wolverhampton,
United Kingdom

[2]TALN, Department of Information and Communication Technology, Universitat Pompeu Fabra,
Spain

sanjastajner@wlv.ac.uk

Abstract. This study addresses the automatic simplification of texts in Spanish in order to make them more accessible to people with cognitive disabilities. A corpus analysis of original and manually simplified news articles was undertaken in order to identify and quantify relevant operations to be implemented in a text simplification system. The articles were further compared at sentence and text level by means of automatic feature extraction and various machine learning classification algorithms, using three different groups of features (POS frequencies, syntactic information, and text complexity measures) with the aim of identifying features that help separate original documents from their simple equivalents. Finally, it was investigated whether these features can be used to decide upon simplification operations to be carried out at the sentence level (split, delete, and reduce). Automatic classification of original sentences into those to be kept and those to be eliminated outperformed the classification that was previously conducted on the same corpus. Kept sentences were further classified into those to be split or significantly reduced in length and those to be left largely unchanged, with the overall F-measure up to 0.92. Both experiments were conducted and compared on two different sets of features: all features and the best subset returned by an attribute selection algorithm.

Keywords. Spanish text simplification, supervised learning, sentence classification.

Eliminación de frases y decisiones de división basadas en corpus para simplificación de textos en español

Resumen. Este estudio aborda el problema de simplificación automática de textos en español con el fin de hacerlos más accesible a las personas con discapacidades cognitivas. Análisis de corpus de artículos originales y artículos simplificados manualmente se ha realizado para identificar y calificar relevantes operaciones que tienen que ser implementadas en el sistema de simplificación de textos. Luego los artículos se han comparado al nivel de frase y texto mediante extracción automática de características y diversos algoritmos de aprendizaje de máquina para clasificación usando tres distintos grupos de características (frecuencias de partes de oración (POS), información sintáctica y medidas de la complejidad de textos) con el propósito de identificar las características que ayuden a distinguir los documentos originales de sus simples equivalentes. Finalmente, se ha investigado la posibilidad de usar esas características en operaciones de simplificación a nivel de frase (dividir, eliminar y reducir). Clasificación automática de frases originales en las que deben preservarse y las que deben eliminarse ha superado la clasificación anterior sobre el mismo corpus. Las frases guardadas luego se clasificaron en las que se dividen o reducen de manera significativa en su longitud y las que se quedan sin cambios mayores con la F-medida de 0.92. Ambos experimentos se realizaron y compararon sobre dos distintos conjuntos de características: el de todas características y el mejor subconjunto recuperado por el algoritmo de selección de atributos.

Palabras clave. Simplificación de textos en español, aprendizaje supervisado, clasificación de frases.

1 Introduction

With the emergence of organisations that promote universal web accessibility in order to enable inclusion of all people into the cultural life of their communities, automatic text simplification as an natural language processing (NLP) task takes the stage. Given that a lot of online content comes in the form of written text, it is important that this text be presented in a particular manner so as to ensure that certain groups of users, such as people with cognitive disabilities, can make productive use of it. According to European easy-to-read guidelines [23], this is mainly achieved by expressing a single idea per sentence, while WCAG2.0 guidelines [16] indicate that most commonly used words in a language should substitute any complex words or phrases in easy-to-read texts. These are but a few of the numerous easy-to-read guidelines, put forward by organisations such as Web Accessibility Initiative[1]. There have been attempts to manually create simplified reading material based on an "original", as is the case with the well-known Simple English Wikipedia[2]. However, it is clear that no manual simplification can match the rate of production of new written material on the Web. Hence the introduction of automatic text simplification systems, that operate on various linguistic levels of the text - from syntactic simplification, to synonym substitution, content elimination, and insertion of definitions of difficult terms and concepts.

Several existing parallel corpora of manually simplified texts are used to determine the necessary transformations when simplifying texts in English language: (1) for children [7], using Encyclopedia Britannica and Britannica Elemental [5]; (2) for language learners [33], using original and abridged texts from Literacyworks[3]; (3) for audiences with various reading difficulties [8, 41, 18, 17], using original and Simple English Wikipedia. The analysis of the simplification operations applied by human editors when simplifying texts for language learners in English [33] reported that 30% of sentences were completely eliminated, while 19% of sentences were split into two or more sentences. Another study conducted on a parallel corpus of original and manually simplified texts for people with low literacy levels in Brazilian Portuguese [13] identified sentence splitting as the second most frequent simplification operation, present in 34% of the original sentences (second to lexical substitution present in 46% of the sentences), while it reported that only 0.28% of sentences were completely eliminated.

To the best of our knowledge, no similar analysis exists for text simplification in Spanish, probably due to the lack of such parallel corpora. This study aims to fill that gap by assessing the significance of different simplification operations in a corpus of original and manually simplified Spanish texts for people with cognitive disabilities. Motivated by the studies of Petersen and Ostendorf [33] for text simplification in English, and Gasperin et al. [24] for text simplification in Brazilian Portuguese, this article analyses the types of applied manual transformations in Spanish text simplification and proposes algorithms for classification of original sentences into those which should be *eliminated*, *kept*, *split*, and left *largely unchanged*, as an initial step of building an automatic text simplification system.

The remainder of the article is structured as follows: Section 2 presents an overview of the most relevant previous work in the field of automatic text simplification, and more specifically the automatic classification of the sentences to be simplified; Section 3 describes our approach to the task, the corpus and features, and provides details about the experimental settings; Section 4 discusses the results of the conducted experiments; while Section 5 concludes the article by summarising the main contributions and proposing possible directions for future work.

2 Related Work

Automatic text simplification is either used as a preprocessing tool for other NLP applications [14, 29], or as a tool in its own right, which offers simpler reading material for a variety of target users, such as low literacy individuals [38], readers with aphasia [20], foreign language learners [11], people with cognitive disabilities [22], etc. There have been systems developed for English [32], Portuguese [1], Japanese [28] and Spanish [36], and more recently some work has been done for Basque [4] and Swedish [35]. The earliest text simplification systems were rule-based and focused on syntactic transformations [14, 37], soon followed by works that suggested an additional lexical simplification module, often based on substitution of difficult words with their easier synonyms extracted from WordNet [12]. The principal criterion of word difficulty is word frequency, extracted from the Oxford Psycholinguistic Database [34]. Lal and Ruger [30] and Burstein et al. [11] follow this pattern, while Bautista et al. [6] use a similar approach

[1] http://www.w3.org/WAI/
[2] http://simple.wikipedia.org/wiki/Main Page
[3] http://literacynet.org/cnnsf/index_cnnsf.html

but also factor in word length when determining how difficult a given word is. With De Belder et al. [19], attention is drawn to the problem of word sense disambiguation, since a large number of words, especially the more frequent ones, tend to be polysemic. A recently developed lexical simplification system for Spanish – LexSiS [9] – uses a word vector model, word frequency, and word length to find the most suitable word substitute. It relies on freely available resources, such as an online dictionary and the Web as a corpus.

As a result of the availability of large parallel corpora for English, text simplification has become more data-driven in recent years. Biran et al. [8] and Yatskar et al. [41] apply an unsupervised method for learning pairs of complex and simple synonyms from a corpus of texts from the original Wikipedia and Simple English Wikipedia. Coster and Kauchak [18, 17], approach the problem of text simplification as an English-to-English translation problem using the parallel corpus of aligned sentences from original and simple English Wikipedia. Although this approach might work well for text simplification in English (offering a large parallel corpus and thus enabling machine learning oriented text simplification), it cannot be extended to other languages, as Simple Wikipedia is available only in English. Another limitation of this approach is that, although it employs Basic English vocabulary, shorter sentences and simpler grammar (according to its guidelines), Simple English Wikipedia does not follow easy-to-read guidelines for writing for people with cognitive disabilities and, therefore, does not represent good training material for text simplification for this target audience.

Several previous studies tackled the issue of classification of original sentences into: (1) those to be eliminated and those to be kept, and (2) those to be split and those to be left unsplit, as an initial step of an automatic text simplification system. In terms of the Spanish language, Drndarevic and Saggion [21] obtained an F-measure of 0.79 for the first classification problem (1), using the SVM classifier on the same data set. For the English language, Petersen and Ostendorf [33] reported an average error rate of 29% for the second classification (2) using the C4.5 decision tree learner, while Gasperin et al. [24] achieved an F-measure of 0.80 using the SVM classifier for the same task in Brazilian Portuguese. Experiments presented in this study outperformed previous results on both classification tasks.

3 Methodology

This study consists of three main parts: (1) a corpus study which quantitatively analyses and categorises simplification operations applied by human editors with the aim of measuring the impact of this operations in an automatic text simplification system; (2) a comparison of original (O) and simplified (S) texts using three different groups of features (POS tags, syntactic features, and complexity measures) with the aim of verifying whether original and simplified texts can be separated automatically according to these features; and (3) two sentence classification experiments in order to explore whether the target sentences for some of the operations found in (1) could be automatically selected using the features and findings from (2).

3.1 Corpus and Features

The study is based on a corpus of 37 pairs of original news articles in Spanish (published online and obtained from the news agency Servimedia[4]), and their manual simplifications, obtained by following easy-to-read guidelines for people with cognitive disabilities. The corresponding pairs of original and manually simplified texts were automatically aligned at sentence level using a tool created for this purpose [10], upon which alignment errors were corrected manually. All texts were further parsed with state-of-the-art Connexor's Machinese syntax parser[5], and 29 features (Table 1) were automatically extracted using the parser's output.

Three sets of features were considered: POS frequencies, syntactical features, and text complexity features (Table 1). The use of the first and second set of features was inspired by the syntactic concept of the projection principle [15] used in [39], and by studies of Petersen and Ostendorf [33], and Gasperin et al. [24]. The features in the third group can be seen as different text complexity indicators: the first three – *scom*, *asl*, and *sci* – refer to the sentence and syntactic complexity; while the other three – *awl*, *ld*, and *lr*

[4] http://www.servimedia.es/

[5] www.connexor.eu

Table 1. Features

Group	Code	Feature
(I) POS tags	*v*	verb
	ind	indicative
	sub	subjunctive
	imp	imperative
	inf	infinitive
	pcp	participle
	ger	gerund
	adj	adjective
	adv	adverb
	pron	pronoun
	det	determiner
	n	noun
	prep	preposition
	cc	coordinate conjunction
	cs	subordinate conjunction
(II) Syntactic	*main*	head of the verb phrase
	premark	preposed marker
	premod	pre-modifier
	postmod	post-modifier
	nh	head of the noun phrase
	advl	head of the adverbial phrase
(III) Complexity	*scom*	simple vs. complex sentences ratio
	asl	average sentence length
	sci	sentence complexity index
	awl	average word length
	ld	lexical density
	lr	lexical richness
	punc	average number of punctuation marks
	num	average number of numerical expressions

– represent lexical complexity. *Sci* was used as a readability measure in second language learning studies [2, 3], while *scom* was used as a complexity feature in [40]. Six of these complexity features (all except *ld* and *lr*) were used by Drndarevic et al. [22] for the evaluation of the degree of simplification in an automatic text simplification for Spanish.

3.2 Experimental Settings

The study consisted of four main experiments:

- Corpus study of simplification operations applied by human editors (Section 4.1);
- Analysis of differences between original and simplified texts, and the corresponding classification experiments (Section 4.2);
- Sentence classification into those to be eliminated and those to be kept during the simplification process (Section 4.3);
- Sentence classification into those to be split and those to be kept largely unchanged (Section 4.4).

Analysis of differences between original and simplified texts was based on all three groups of features (Table 1), while the corresponding classification experiment (Section 4.2) used only the third group of features (complexity measures). The two sentence classification experiments (Sections 4.3 and 4.4) were based only on the first two groups of features (POS tags and syntactic features) and two additional features – *words* (sentence length in words), and *sent* (position of the sentence in the text). All classifications were performed in Weka[6] [27, 25], with the 10 cross-fold validation setup. In the third and fourth experiments (Sections 4.3 and 4.4), first the CfsSubsetEval attribute selection algorithm [26] implemented in Weka was used to select a subset of best features, after which all classification algorithms were applied to both – the whole feature set (all), and to the 'best' features returned by the CfsSubsetEval algorithm (best).

4 Results and Discussion

In the first step of the analysis, simplification operations applied by human editors were selected, categorised and analysed quantitatively (Section 4.1). Subsequently, the differences between original and simplified texts based on the three groups of features (Table 1) were analysed with additional classification of texts into original and simple ones, based only on the third group of the features (Section 4.2). Finally, classification algorithms were derived for the classification of 'original' sentences into the ones to be deleted vs. the ones to be kept (Section 4.3); and into those to be split vs. those to be left largely unchanged (Section 4.4).

4.1 Corpus Study

A total of 468 manually applied simplification operations were observed in the corpus. Out of a total of 247 original sentences, 44 were left unchanged (18%), largely in the case of headlines. Most often, one operation type was applied per sentence (46%), but as many as seven different operations were applied to a single original sentence. It is important to point out that certain operations, such as summarisation and paraphrases, even though they count as a single operation, often involve complex restructuring of the original sentence, as in the following pair of original (1) and simplified (2) sentences[7]:

1. *"The organisation that bears the name of the founder of the publishing house SM, assured in a press release that the centre will mark the inauguration of a new culturual, educational and social project, which, by placing literature and reading in the spotlight, aims at opening the door to new possibilities that these two bring to our present moment."*
2. *"The Reading Centre is a cultural, educational and social project. Reading and the world of books will take the leading role in this project."*

The observed transformations have been classified into three broad categories: (1) lexical transformations, which include substitution of an original word with a synonym or a near synonym, and explanation of metaphorically used

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] All examples in the article are translated into English so as to make it more legible.

Table 2. Analysis of original and simplified texts: POS features – Group I (on average per sentence – (s), and on average per text – (t))

Corpus	*v*	*ind*	*sub*	*imp*	*inf*	*pcp*	*ger*	*adj*	*adv*	*pron*	*det*	*n*	*prep*	*cc*	*cs*
Original (s)	2.75	1.73	0.12	0.00	0.46	0.39	0.05	1.85	0.61	1.40	3.93	9.42	5.53	0.86	0.38
Simplified (s)	1.78	1.24	0.07	0.00	0.18	0.28	0.00	0.61	0.33	0.57	2.00	4.55	2.13	0.26	0.20
Original (t)	19.78	12.38	0.89	0.03	3.30	2.81	0.38	13.22	4.57	9.78	27.57	66.22	38.73	6.00	2.81
Simplified (t)	16.00	11.11	0.65	0.00	1.70	2.49	0.05	5.22	3.03	5.03	17.38	39.59	18.68	2.30	1.76

Table 3. Analysis of original and simplified texts: Complexity features – Group III (on average per text)

Corpus	*punc*	*num*	*word*	*sent*	*ASL*	*AWL*	*scom*	*sci*	*ld*	*lr*
Original	16.41	3.54	191.41	7.21	27.10	5.14	2.70	13.88	0.48	0.46
Simple	8.95	1.43	109.05	8.89	12.44	4.88	1.11	6.43	0.46	0.43

expressions; (2) syntactic transformations, which include sentence splitting, sentence deletion, and turning passive constructions into active ones; and (3) transformations applied to the lexical and syntactic levels simultaneously, which include paraphrases and summarisation. The latter category accounts for the majority of transformations (44%), whereas lexical and syntactic operations are relatively equally present (22% and 25% respectively). Syntactic operations are also illustrated in the variation of the number of clauses (identified by finite verb forms) between original and simplified sentences. Even though in a large number of cases (39% of all sentence pairs) the number of clauses is lower in simplified sentence (as a result of summarisation), in 47 pairs the number of clauses actually increased when simplifying the original sentence. This is due to expanding non-finite verb phrases into finite predicates, easier to understand for people with cognitive disabilities, as in the following example:

1. *“Ban Ki-moon: More humanitarian aid is needed.”*
2. *“Ban Ki-Moon says that more humanitarian aid is needed.”*

It is worth noting that the expansion is perceived at another level – there are 54 simplified sentences without an original counterpart. All of these are cases of definitions of difficult terms and concepts, introduced when no synonym substitution is feasible. Thus, for example, *Amnesty International* is defined as “an organisation that helps people around the World" in simplified texts.

Observing the transformations on the sentence level, it is interesting to notice that a quarter of all original sentences were split into two (71% of all split sentences) or more (three or four, 22% and 7% respectively) simplified ones. However, further analysis of the unsplit sentences (those with ‘1-1’ alignment) showed that sentence length was significantly reduced in some of the simplified sentences with respect to their original counterpart, as in the following pair of original (1) and simplified (2) sentences:

1. *“Matute (Barcelona, 1926) was seen as the main contender for the Cervantes award this year, and after the last year’s victory by the Mexican author José Emilio Pacheco, the unwritten rule is confirmed, by which an author from Latin America and another one from Spain take turns to win the award."*
2. *“The Mexican author Emilio Pacheco won the award last year.”*

Therefore, the unsplit sentences were further grouped into those where the sentence length was reduced by more than 10 words – *reduced*, and those where the change in sentence length was less than 10 words – *same*. This analysis showed that out of 140 unsplit sentences, 66 sentences (26% of all sentences) had their length significantly reduced during the simplification process, while 74 sentences (29% of all sentences) had their length either left unchanged or slightly changed.

Based on their transformation type, all sentences of the original set could be classified into four groups: *deleted*, *split*, *reduced*, and *same*. It is worth noting that in the used corpus, these four groups were almost equally distributed: *deleted* (21%), *split* (23%), *reduced* (26%), and *same*

(29%). Automatic detection and classification of all original sentences into these four groups would, therefore, be an important step in automatic text simplification.

4.2 Differences between Original and Simplified Texts

The frequencies of the selected features (Table 1), in the original (O) and simplified (S) texts, were calculated as an average per sentence (only the features belonging to the first two groups) and per text (all features). The results are presented in Tables **??**, 4, and 3. Calculated as an average per text, most of the features reported a statistically significant difference at a 0.002 level of significance (paired t-test implemented in SPSS), with the exceptions being *ind* (0.035), *ger* (0.003), *ld* (0.039), *lr* (0.004), and the three features which did not report any statistically significant difference: *sub*, *imp*, and *pcp*.

One could argue that complexity features (group III) could be seen as a kind of readability or complexity measure. Therefore, it would be interesting to see if any (or any subset) of them could be used to automatically estimate whether a text is simple enough to be understood by people with cognitive disabilities. In order to test this assumption, the rule-based JRip classification algorithm was used for the classification into complex (original) and simplified texts. First, the classification was performed using all eight features (*punc*, *num*, *asl*, *awl*, *scom*, *sci*, *ld*, *lr*). As the JRip algorithm returned the rule that used only one feature – *asl*, with excellent performance (F measure = 0.99), in the next step, the *asl* was excluded and the JRip classification algorithm was applied again. The returned rule used only the *sci* feature and had the same performance as in the previous case (Table 5). In the subsequent step, both *asl* and *sci* were excluded, and the JRip algorithm returned the rule that used *punct* and *awl*, this time with slightly lower performance than in the previous cases (Table 5). With each new step from this point on (excluding all features returned by JRip rules so far), the performance of the classification algorithms drastically decreased. Two additional classification algorithms – a tree based J48 algorithm and an SVM algorithm with standardisation, were used to ensure that the drop in the classification performance is not caused by the choice of the algorithm but rather by the set of features used. These findings (summarised in Table 5) indicated that the syntactic and structural transformations (e.g. sentence splitting, sentence reduction, etc.) which modify the average sentence length are a key operation when simplifying texts for people with cognitive disabilities.

4.3 Sentence Elimination

The analysis of sentence transformation (Section 4.1) showed that 21% of the sentences were eliminated in the process of manual simplification. Therefore, automatic detection of sentences to be deleted would be an important step in automatic text simplification. This problem was already addressed in [21], using the same data set but different features and classification algorithms. In the said study, the authors borrowed features from text summarisation and added new ones (e.g. position of the sentence in the text, and number of named entities, numerical expressions, content words and punctuation tokens). Their classification system, based on an SVM implementation [31], outperformed both baselines: the one deleting the last sentence, and the one deleting last two sentences in each document. In this study, we used the first and second group of features (Table 1), and two additional features: the position of the sentence in the text (*sent*) and number of words in the sentence (*words*). The sentences were classified into *deleted* and *kept*, using the JRip and J48 classifiers, on both – the entire set of features (all), and only the subset of best features (best) returned by the CfsSubsetEval attribute selection algorithm. Both algorithms (in both feature set-ups) outperformed the SVM classifier in [21] (Table 6). The greatest improvements were achieved in terms of precision in classifying *deleted* sentences (P = 0.85 for JRip(best)) and recall in classifying *kept* sentences (R = 0.99 for JRip(best) and J48(best)). Figure 1 presents the rules returned by the JRip classifier when using all features (all), and only the 'best' subset of initial features – *sent*, *noun*, *words* (best).

4.4 Sentence Splitting

After classifying sentences of the original set into those to be kept and those to be eliminated, the next step is the classification of kept sentences into the ones to be split or significantly reduced in length (*split* + *reduced*), and the ones to

Table 4. Analysis of original (O) and simplified (S) texts: Syntactic features – Group II (on average per sentence – (s), and on average per text – (t))

Corpus	*main*	*premark*	*premod*	*postmod*	*nh*	*advl*
Original (s)	2.48	3.06	5.90	4.28	9.39	0.50
Simple (s)	1.52	1.12	2.89	1.67	4.48	0.24
Original (t)	17.78	21.79	41.62	30.03	65.97	3.68
Simple (t)	13.73	9.89	24.95	14.38	39.19	2.30

Table 5. Classification into original and simplified texts using complexity features

Feature set	**JRip**	**J48**	**SVM**
{*punc*, *num*, **asl**, *awl*, *scom*, *sci*, *ld*, *lr*}	0.99	0.99	1
{*punc*, *num*, *awl*, *scom*, **sci**, *ld*, *lr*}	0.99	0.99	1
{**punc**, *num*, **awl**, *scom*, *ld*, *lr*}	0.95	0.95	0.96
{*num*, **scom**, *ld*, *lr*}	0.73	0.73	0.71
{**num**, *ld*, *lr*}	0.62	0.55	0.67
{*ld*, **lr**}	0.57	0.46	0.67
{*ld*}	0.50	0.40	0.58

Rules returned by JRip algorithm for both feature sets – (all) and (best):

```
(sent >= 4) and (noun <= 4) => sent_type = deleted
=> sent_type = kept
```

Fig. 1. *Deleted* vs. *kept* sentences

be left practically unchanged (*same*). In this classification, the *reduced* sentences (in the sense explained in Section 4.1) were treated as split, as they could be seen as sentences which were split into a significantly shorter sentence that was kept and another that was eliminated[8]. The CfsSubsetEval attribute selection algorithm returned {*noun*, *premod*, *nh*, *advl*, *words*} as the best subset of features. The results of this classification (using both the full set of features and the 'best' features only) are presented in Table 7, and the rules returned by the JRip classifier are shown in Figure 2.

In this classification, the SVM algorithm implemented in Weka (SMO) was used in order to investigate whether the JRip and J48 algorithms perform significantly worse than SVM for this type of classification. SMO was used in all three set-ups: with standardisation (-s); with normalisation (-n); and without normalisation or standardisation. The statistical significance of differences between the results was measured by paired t-test, offered in Weka Experimenter. None of the algorithms reported significantly better performance than any other (at a 0.05 level of significance), nor did any feature set outperform any other.

Rules returned by JRip algorithm for both feature sets – (all) and (best):

```
(words <= 18) => sent_type=same
 => sent_type=split
```

Fig. 2. *Split* vs. *same* sentences

5 Conclusions and Future Work

This study presented a thorough analysis of a parallel corpus of original and manually

[8] An additional experiment of classifying the original sentences into the ones to be *split* or significantly *reduced* showed that these two groups are very similar. None of the five classification algorithms (with both feature sets – 'all' and 'best') achieved the F measure higher than 0.66 (the best being J48-all). This can be seen as an additional proof that the *split* and *reduced* sentences should be treated as belonging to the same group.

Table 6. Classification into *deleted* and *kept* sentences

Method	Deleted			Kept			Overall
	P	R	F	P	R	F	F
Delete last	0.27	0.20	0.23	0.81	0.86	0.84	0.73
Delete 2 last	0.31	0.46	0.37	0.84	0.74	0.79	0.68
SVM [21]	0.42	0.26	0.30	0.86	0.89	0.87	0.79
JRip(all)	0.81	0.32	0.46	0.84	0.98	0.91	0.81
JRip(best)	**0.85**	0.32	0.47	0.84	**0.99**	0.91	**0.82**
J48(all)	0.45	0.45	0.45	0.85	0.85	0.85	0.77
J48(best)	**0.83**	0.28	0.42	0.84	**0.99**	0.91	**0.80**

Table 7. Classification into *split* and *same* sentences

Method	Split			Same			Overall
	P	R	F	P	R	F	F
JRip(all)	0.93	0.93	0.93	0.88	0.88	0.88	0.91
JRip(best)	0.93	0.91	0.92	0.85	0.88	0.87	0.90
J48(all)	0.92	0.88	0.90	0.81	0.88	0.84	0.88
J48(best)	0.93	0.89	0.91	0.82	0.88	0.85	0.89
SMO-n(all)	0.94	0.90	0.92	0.85	0.90	0.87	0.90
SMO-n(best)	0.97	0.89	0.93	0.83	0.96	0.89	0.91
SMO-s(all)	0.93	0.90	0.91	0.83	0.88	0.85	0.89
SMO-s(best)	0.95	0.91	0.93	0.86	0.92	0.89	0.91
SMO(all)	0.90	0.91	0.91	0.85	0.84	0.84	0.88
SMO(best)	0.95	0.92	0.93	0.87	0.92	0.89	0.92

simplified texts for people with cognitive disabilities. It showed that original and simplified texts significantly differ on all 29 investigated features that belong to three different groups (POS tags, syntactic features and complexity measures). The analysis of manual transformations revealed that in most cases only one operation type was applied per sentence, but as many as seven different operations were applied to a single original sentence in some cases. It also showed that summarisation and paraphrases were the most prominent operations (44%), whereas lexical and syntactic operations were relatively equally present (22% and 25%, respectively). The classification between original and simplified texts using only the third group of features (complexity measures) indicated that the transformations that modify the average sentence length (e.g. sentence splitting, sentence reduction, etc.) are crucial when simplifying texts for people with cognitive disabilities. All “original" sentences were further divided into four groups: *deleted*, *split*, *reduced*, and *same* (*same* in the sense that they are either left unchanged or the number of words in the original and its corresponding simplified sentence did not differ for more than 10 words), suggesting automatic classification of original sentences into these four groups as an initial step in automatic text simplification. The first classification task, between *deleted* and *kept* (*split*, *reduced*, and *same*), outperformed the systems previously applied on the same corpus (achieving the F-measure of up to 0.82), and returned the sentence position in the text (*sent*) and number of nouns in the sentence (*noun*) as the most important features. Further classification of the *kept* sentences into the *split* (actually containing both groups *split* and *reduced*) ones and the *same* ones, achieved the F-measure of up to 0.92.

We are currently working on the alignment of an additional 163 manually simplified texts and their originals in order to enlarge the training set for the two classification tasks. After that, the goal is to implement an automatic classification system which would classify original sentences into the four aforementioned groups. Finally, we will analyse the process of manual sentence splitting in order to build an automatic simplification system and apply it to the sentences marked as the ones to be split (by the classification algorithm).

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 [communication] 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. We acknowledge partial support from the following grants: Avanza Competitiveness grant number TSI-020302-2010-84 from the Ministry of Industry, Tourism and Trade, Spain and grant number TIN2012-38584-C06-03 and fellowship RYC-2009-04291 (Programa Ramón y Cajal 2009) from the Spanish Ministry of Economy and Competitiveness.

References

1. **Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E., & De Mattos Fortes, R. P.** (**2008**). Towards Brazilian Portuguese automatic text simplification systems. In *ACM Symposium on Document Engineering*. 240–248.
2. **Anula, A.** (**2007**). Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*. 45–61.
3. **Anula, A.** (**2008**). Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. In *La evaluación en el aprendizaje y la enseñanza del español como LE/L2, Pastor y Roca (eds.)*. Alicante, 162–170.
4. **Aranzabe, M. J., Díaz De Ilarraza, A., & González, I.** (**2012**). First Approach to Automatic Text Simplification in Basque. In *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop at LREC 2012*.
5. **Barzilay, R. & Elhadad, N.** (**2003**). Sentence alignment for monolingual comparable corpora. In *Proceedings of the EMNLP conference*.
6. **Bautista, S., Gervás, P., & Madrid, R.** (**2009**). Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In *The Second International Conference on Information and Communication Technologies and Accessibility*.
7. **Bautista, S., Len, C., Hervs, R., & Gervs, P.** (**2011**). Empirical identification of text simplification strategies for reading-impaired people. In *Proceedings of the European Conference for the Advancement of Assistive Technology.*
8. **Biran, O., Brody, S., & Elhadad, N.** (**2011**). Putting it Simply: a Context-Aware Approach to Lexical Simplificaion. In *Proceedings of the ACL*.
9. **Bott, S., Rello, L., Drndarevic, B., & Saggion, H.** (**2012**). Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), Mumbai, India, 8-15 December.*
10. **Bott, S. & Saggion, H.** (**2011**). Spanish Text Simplification: An Exploratory Study. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47.
11. **Burstein, J., Shore, J., Sabatini, J., Lee, Y.-W., & Ventura, M.** (**2007**). The Automated Text Adaptation Tool. In *HLT-NAACL (Demonstrations)*. 3–4.
12. **Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J.** (**1998**). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. 7–10.
13. **Caseli, H., Pereira, T., Specia, L., Pardo, T., Gasperin, C., & Aluísio, S.** (**2009**). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Proceedings of the 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), March 01–07, Mexico City*.
14. **Chandrasekar, R., Doran, D., & Srinivas, B.** (**1996**). Motivations and Methods for Text Simplification. In *Proceedings of COLING*. 1041–1044.
15. **Chomsky, N.** (**1986**). *Knowledge of language: its nature, origin, and use.* Greenwood Publishing Group, Santa Barbara, California.
16. **Cooper, M., Reid, L., Vanderheiden, G., & Caldwell, B.** (**2010**). Understanding wcag 2.0. a guide to understanding and implementing web content accessibility guidelines 2.0. World Wide Web Consortium (W3C).

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

18. **Coster, W. & Kauchak, D.** (**2011**). 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*. 665–669.

19. **De Belder, J., Deschacht, K., & Moens, M.-F.** (**2010**). Lexical simplification. In *Proceedings of the 1st International Conference on Interdisciplinary Research on Technology, Education and Communication (LTEC 2010)*.

20. **Devlin, S. & Unthank, G.** (**2006**). Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06. New York, NY, USA, 225–226.

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

22. **Drndarevic, B., Štajner, S., Bott, S., Bautista, S., & Saggion, H.** (**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. Lecture Notes in Computer Science. Samos, Greece, 24-30 March, 2013.*

23. **Freyhoff, G., Hess, G., Kerr, L., Menzel, E., Tronbacke, B., & Van Der Veken, K.** (**1998**). Make it Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability; for authors, editors, information providers, translators and other interested persons.

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

25. **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H.** (**2009**). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11, 10–18. ISSN 1931-0145. doi:http://doi.acm.org/10.1145/1656274.1656278.

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

27. **Ian H. Witten, E. F.** (**2005**). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

28. **Inui, K., Fujita, A., Takahashi, T., Iida, R., & Iwakura, T.** (**2003**). Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*. 9–16.

29. **Klebanov, B. B., Knight, K., & Marcu, D.** (**2004**). Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*. 735–747.

30. **Lal, P. & Ruger, S.** (**2002**). Extract-based Summarization with Simplification. In *Proceedings of the ACL 2002 Automatic Summarization / DUC 2002 Workshop*.

31. **Li, Y., Zaragoza, H., Herbrich, R., Shawe-Taylor, J., & Kandola, J.** (**2002**). The Perceptron Algorithm with Uneven Margins. In *Proceedings of the 9th International Conference on Machine Learning (ICML-2002)*. 379–386.

32. **Medero, J. & Ostendorf, M.** (**2011**). Identifying Targets for Syntactic Simplification.

33. **Petersen, S. E. & Ostendorf, M.** (**2007**). Text simplification for language learners: A corpus analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*.

34. **Quinlan, P.** (**1992**). *The Oxford Psycholinguistic Database*. Oxford University Press.

35. **Rybing, J., Smithr, C., & Silvervarg, A.** (**2010**). Towards a Rule Based System for Automatic Simplification of Texts. In *Proceedings of the Third Swedish Language Technology Conference*.

36. **Saggion, H., E., G.-M., Etayo, E., Anula, A., & Bourg, L.** (**2011**). Text Simplification in Simplext: Making Text More Accessible. *SEPLN Journal*, 47, 341–342.

37. **Siddharthan, A.** (**2002**). An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC 2002)*. 64–71.

38. **Specia, L.** (**2010**). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*. Berlin, Heidelberg. ISBN 3-642-12319-8, 978-3-642-12319-1, 30–39.

39. **Štajner, S., Evans, R., Orasan, C., & Mitkov, R.** (**2012**). What Can Readability Measures Really Tell

Us About Text Complexity? In *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. Istanbul, Turkey. ISBN 978-2-9517408-7-7.

40. **Štajner, S. & Mitkov, R.** (**2012**). Diachronic Changes in Text Complexity in 20th Century English Language: An NLP Approach. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.

41. **Yatskar, M., B., P., Danescu-Niculescu-Mizil, C., & Lee, L.** (**2010**). For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*. 365–368.

Sanja Štajner is currently a second year PhD student at the University of Wolverhampton (UK), working on text simplification for people with disabilities. Sanja obtained her B.Sc. in Mathematics and Computer Science at the University of Belgrade (Serbia), and MA in Natural Language Processing and Human Language Technologies from Universitat Autonoma de Barcelona (Spain) and the University of Wolverhampton (UK). Her main research interests include natural language processing, machine learning, statistical analysis and corpus linguistics.

Biljana Drndarević obtained her master degree in Natural Language Processing in June 2011 and was awarded a joint degree from the University of Franche-Comté and the Autonomous University of Barcelona. Since 2012 she has worked on the Simplext project for automatic text simplification at Pompeu Fabra University in Barcelona.

Horacio Saggion was born in Campana, Buenos Aires, Argentina. He holds a PhD in Computer Science from Universite de Montreal, Canada. He obtained his BSc in Computer Science from Universidad de Buenos Aires in Argentina, and his MSc in Computer Science from UNICAMP in Brazil. Horacio is currently a Ramón y Cajal Research Professor at the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona. He is associated with the Natural Language Processing group where he works on automatic text summarisation, text simplification, information extraction, sentiment analysis and related topics. His research is empirical, combining symbolic, pattern-based approaches and statistical and machine learning techniques. Horacio has published over 70 works in leading scientific journals, conferences, and books in the field of human language technology.

Article received on 13/12/2012; accepted on 25/01/2013.