Syntactic Simplification for Machine Translation

José CAMACHO COLLADOS

Centre Tesnière, Université de Franche-Comté, France
Research Group in Computational Linguistics, University of Wolverhampton, United Kingdom
Laboratori fLexSem, Universitat Autònoma de Barcelona, España

Abstract

This paper presents a new rule-based Syntactic Simplification system for the Spanish language. The system basically consists of splitting complex sentences into several simple ones preserving its original meaning. It is envisioned to be a preprocessing tool for other Natural Language Processing applications such as Text Summarization, Information Extraction, Parsing and Machine Translation. According to the evaluation, the application of the syntactic rules proposed clearly improves Machine Translation from Spanish to far target languages such as Korean and Chinese. A parallel corpus of Spanish original and simplified texts has been manually built and validated by native Spanish speakers for this research. In addition, an experiment was carried out among two groups of Spanish learners comparing their reading comprehension of original and simplified texts. Their answers show a null correlation between simplification and text comprehension. The automatic Syntactic Simplification system has also been intrinsically evaluated obtaining promising results.

Key-words

Text Simplification; Syntactic Simplification; Natural Language Processing; Spanish; Machine Translation.

1. Introduction

Text Simplification (TS) is a Natural Language Processing (NLP) task consisting of reducing the complexity of a sentence in order to make it simpler, but without changing the general meaning of it (Siddharthan, 2002: 1). This research focuses on the syntactic level. The syntactic simplifications

considered are basically based on splitting a complex sentence into several simpler ones without damaging its original meaning. The motivations for Text Simplification are basically divided in two main groups. The first one is the one which uses TS in order to make a text accessible for certain people with problems in reading complex structures. This research includes an evaluation of the system regarding this application but, more specifically, for foreigners learning Spanish. The second main group includes the systems aiming to improve other NLP applications, which is the main purpose of this research. Some of the applications which may be improved from this approach are explained below:

– **Text Summarization**: The idea for the TS in Text Summarization is to reduce the information extracted by each sentence, keeping just the relevant one. Since Text Summarization is sometimes based on the information extracted from different sentences, having the information better divided in sentences might be extremely useful for the task (Chandrasekar et al., 1996; Siddharthan, 2002).

– **Information Extraction**: As far as Information Extraction is concerned, automatic systems may work better when the complexity of a sentence is low (Evans 2011, Klebanov et al., 2004). TS may help split different types of information inside a complex sentence into several simple sentences.

– **Parsing**: The goal of TS in parsing is due to the improvement of performance if given a short sentence as input (Siddharthan, 2002). Longer sentences give more cases of ambiguity, so that TS aims to reduce the length of the sentences while preserving the meaning as much as possible.

– **Machine Translation**: One of the evaluations of the Syntactic Simplification (SS) system carried out on this research is extrinsic and it is done regarding the improvement of some MT systems taking Spanish as source language. The target languages taken into account in this research are Chinese and Korean. The far languages (such as Spanish and either Chinese or Korean) and long sentences are nowadays two of the main drawbacks of MT, so it may be quite useful to have a monolingual tool simplifying complex sentences regardless of the target language.

This paper is organized as follows: Section 2 shows previous researches related to Text Simplification and Machine Translation; Section 3 explains SS splitting rules considered along this research; Section 4 introduces the Spanish parallel corpus of original and syntactically simplified sentences and its validation; Section 5 shows the development and architecture of the automatic SS system by applying the rules presented in Section 3; Section includes the both extrinsic and intrinsic evaluations of the system; and Section 7 gives the conclusions of the work and address some possible directions for further research on the subject.

2.Related work

2.1 First steps in Text Simplification

Chandrasekar et al. (1996) was the first serious attempt in order to create a general architecture for a TS system. In the research, they comment the usefulness of the TS for a wide variety of applications. The main applications have been explained in Subchapter 1.1. Apart from helping NLP applications such as MT, parsing and Text Summarization, Chandrasekar et al. (1996) introduce the possibility that TS could help Information Retrieval. Not only are NLP applications addressed as possible beneficiaries of the TS, but also TS could make the text clear and easier to understand. TS may be used as a preliminary step in order to create this artificial constrained language like controlled languages, which will be handled in Sub-section 2.4. Nevertheless, the research was focused on improving the performance of a parser, which consequently would be beneficial for the other applications. Regarding this task of improving a parser, the main problem was to reduce the complex syntactic structures that we can find in long sentences. This was achieved by opportunistically splitting the complex sentences into two or more simple ones.

Siddharthan (2002) proposes a new architecture for a TS system. It follows the work of Chandrasekar et al. (1996) but introducing quite remarkable improvements. The first remarkable difference from the previous work is the introduction of a new stage on the TS process: regeneration. This is the main improvement of this research and where Siddharthan tries to focus on. The discourse level problems raised by Chandrasekar et al. (1996) are partially solved by this new stage in the system. The new architecture is then based in three stages: analysis, transformation and regeneration. As far as SS is

concerned, Siddharthan (2002) considered several syntactic phenomena such us relative clauses, adverbial clauses, coordinated clauses, subordinated clauses, correlated clauses, participial phrases, appositive phrases and passive voice. Afterwards, he analyzes and evaluates each stage separately, giving more emphasis to the new regeneration stage.

2.2 Corpus-based Text Simplification

Several rule-based systems were created after Siddharthan's work. Some of them already gave some external evaluation depending on the purpose of the system (Petersen and Ostendorf, 2007; Evans, 2011). There are also works addressing the need of a parallel corpus of original and simplified sentences (Petersen and Ostendorf, 2007; Aluísio et al., 2008; Specia et al., 2009; Specia, 2010). The corpus, apart from potentially being used to develop a Machine Learning algorithm, would be useful to carry out a deeper analysis of the task leading to some new ideas or improvements of rule-based TS systems.

Specia et al. (2009) followed this direction and worked on building a Brazilian Portuguese parallel corpus of original and simplified sentences with both lexical and syntactic simplifications. The main goal of the corpus is to help people with low level of literacy or some other cognitive disabilities. Specia (2010) experimented afterwards with a quite simple corpus-based TS approach for the Brazilian Portuguese. The goal of this system is the same as in Specia et al. (2009), for people with some problems at reading. She basically used a Statistical Machine Translation (SMT) method to deal with the TS problem and check the results. The SMT was carried out without making many changes, so that the approach could be easily improved by adapting the framework to the particular TS problem.

2.3 Spanish Text Simplification

Bott and Saggion (2011a) studies the problem of TS for the Spanish language. The goal of Bott and Saggion (2011a) is to create a text easy to read for people with learning disabilities, so it differs in this point from the main goal of this research. It is a preliminary study where they analyze a corpus of news and the respective simplified one. They address the need of getting a parallel corpus in order to be able to use it for the creation of a reliable TS system.

Bott and Saggion (2011b) follows their last paper and explain in detail an algorithm to align a parallel corpus of news and their simplified ones at a sentence level. As there is not training data for task (there is no manually aligned parallel corpus), they rely on unsupervised learning.

Drndarević et al. (2013) presents a two-component (syntactic and lexical) automatic text simplification system for the Spanish language in order to make the text easier to read for people with cognitive disabilities. The system managed to get simpler sentences without seriously damaging their grammaticality and preservation of meaning with the original sentences. However, they do not propose any solution to the parsers errors, which caused most of the errors found in the system.

2.4 Controlled Languages and Machine Translation

Cardey et al. (2004) addressed the problem of MT from French to two far languages such as Chinese or Arabic. One of the main conclusions is that many issues concerning a pair of languages should be studied separately for each different language pair, which makes these MT systems really language-independent. This is a problem as many pairs of minority languages will not receive the needed attentions to solve all the specific problems, especially on long sentences. This may be also the case of SMT, since SMT usually backfires when translating long and complex sentences (Koehn, 2010). SMT tends to backfire when the source sentence is long and have several verb phrases, no matter the target language. The problems of MT often come, apart from the length of the sentence, from the ambiguity that such sentence could have. MT systems should deal with this problem, which is one of the main ones for every system. Even deeply treated, ambiguity is really complicate to be totally solved in order to create a high quality MT system. That is one of the main reasons why controlled languages were developed (Kaji, 1999; Mitamura, 1999; Cardey, 2011). Controlled languages are artificially created subset of the natural language where the ambiguity is eliminated and complexity is reduced. This clearly improves MT and it is especially useful where used for different target languages.

Although controlled languages belong to natural language, they are artificially created. They need to be created and people using it should know

all the restrictions in order to be able to use it. As explained in Mitamura (1999), this is not an easy task because usually there is a big restriction in both grammar and vocabulary, so it should be deeply studied before using it. As controlled languages have been proved to be a successful approach for MT on many cases, TS may be used as a tool to reach the desirable controlled language. Temnikova (2012) shows how the post edition by a human translator becomes an easier task by applying TS prior to the MT.

3. Syntactic Simplification splitting rules

The simplification considered in the corpus is basically reduced in splitting long and complex sentences in simple ones. Since this research is aimed to be used to build an automatic simplification system, all the rules have been carefully selected in order to achieve this goal in the future (Camacho Collados, 2013). To begin with, the first condition to split a sentence is the number of conjugated verbs. The sentences which are to be simplified must have at least two conjugated verbs. It is important to notice that in most sentences there are different types of structures to be simplified simultaneously, which makes the task harder. Sub-section 3.1 handles the coordinate sentences and Sub-section 3.2 summarizes some simple subordinate cases.

3.1 Coordination

The original sentence is split on the position of the coordination nexus or articulation point as called by Chandrasekar et al. (1996). It is usually suggested repeating some noun phrases in order to improve the understandability and a best processing by other NLP applications. Example (3) is an illustrative example of how the simplification is handled on coordinate sentences. Several examples on this section are given in English for a better understanding of the reader.

(3) a. Fishes swim in the sea and butterflies fly in the sky.

b. – Fishes swim in the sea.
– Butterflies fly in the sky.

3.2 Subordination

– **Non-restrictive relative clauses:** This kind of structure is split without adding any other element which might cause a mistreatment by other NLP applications. The relative clause begins with a comma followed by a connector such as *que* [which] or *quien* [who]. Example (4) is a simple and representative example for the Spanish language of this type of simplification.

(4) a. Juan, que es aún muy joven, consiguió el premio. [Juan, who is still very young, got the prize.]

b. – Juan es aún muy joven. [Juan is still very young.]
– Juan consiguió el premio. [Juan got the prize.]

– Effect: This type of structure contains the cause-effect relation. They are connected by a conjunction which indicates the end of the cause and the beginning of the effect. Therefore, the splitting of the sentence will be done at the conjunction's position. Another effect connector such as therefore in English is introduced at the beginning of the second sentence, as we can appreciate in Example (5). In Spanish the connector introduced will be *"Por lo tanto,"*.

(5) a. The cat ate poisoned food, so it died.

b. – The cat ate poisoned food.
– Therefore, the cat died.

– Causal: The same kind of relation cause-effect appears in this structure. However, the cause is placed after the effect in this case. They are connected by a different causal connector such as *because* on the English language (*porque* on the Spanish language) - Example (6). The output sentences include the effect in the first position and the cause or reason in the second sentence.

(6) a. Dogs can't fly because they don't have wings.

b. – Dogs can't fly.
– Reason: They don't have wings.

– Indirect speech to direct speech: This structure is currently simplified just on very specific cases. More precisely, in the cases where a

communication verb such as *decir, comunicar or explicar* are followed by the relative que and a sentence concerning the communication of the speaker. Most of the communication verbs included for this task were extracted from Diccionario combinatorio del español contemporáneo (Bosque, 2004). The original sentence is split into two sentences, the second one introduced by *":"*, as we can observe in Example (7):

(7) a. El jugador dijo que el presidente estuvo con el equipo antes del partido. [The player said that the president was with the team before the match.]

b. – El jugador dijo: [The player said:]
– El presidente estuvo con el equipo antes del partido. [The president was with the team before the match.]

4.Corpus

4.1 Creation of the parallel corpus

The corpus chosen as a reference was the AnCora Corpus (Taulé et al., 2008), which consists of Spanish newspaper texts annotated at syntactic and morphological level. Newspaper texts are quite representative of the natural language and complex enough for the simplification task. Once the corpus was transferred from the original XML format to text format, the sentences from the AnCora Corpus were manually simplified as explained in Section 3 in separate text files. On this way a parallel corpus of original and simplified texts was built. The original sentences and the simplified ones are easily aligned, as each original sentence is separated from each other on a new line, which is respected on the simplified part no matter how many new sentences have been created. The corpus currently counts with 3000 original sentences and their respective simplified ones (Camacho Collados, 2013). It is already available for research purposes if required.

4.2 Corpus validation

The validation of the corpus has been done regarding two issues (grammaticality and preservation of meaning), similar to the one used in Drndarević et al. (2013) for the evaluation of their SS system. To carry out this, some preliminary results were obtained from six native Spanish speakers (three of them linguists and three of them holding a non-related

university degree). They required to fill three excel sheets. The first one was to evaluate the grammaticality of the original sentences from AnCora corpus; the second one about the grammaticality of the simplified sentences from AnCora Corpus; and the third one concerning the preservation of meaning of original and simplified sentences. To do so, thirty complex sentences were randomly selected from thirty different texts of the corpus.

All the evaluations were done on a 1-5 scale. For the grammaticality measure, 1 means that sentence is completely a grammatical and the 5 that the sentence is completely grammatical. For the preservation of meaning, 1 means that there is no preservation of meaning at all and 5 that the meaning of the simplified sentence and the original is identical. Table 1 shows the total average for each evaluation and from the results we can appreciate how grammaticality is not damaged on the simplified sentences (4.74 without simplification – 4.66 with simplification). A paired two-sample t-test at the 0.05 level suggests that the difference is not statistically significant ($t(179)=1.513$; p-value=0.132). There are even a few cases where the simplification actually improves the grammaticality of the original sentences.

Table 1: Corpus validation (grammaticality and preservation of meaning)

	Grammaticality Original Sentences	Grammaticality Simplified Sentences	Preservation of meaning
Average	4.74	4.66	4.8
Positive (4-5)	95.6%	97.2%	98.9%
Neutral (3)	3.3%	2.8%	1.1%
Negative (1-2)	1.1%	0%	0%

Regarding the inter-rater reliability for the preservation of meaning task, a statistical test at a 0.05 level was carried out among the six annotators. The results ($F(5,174)=1.577$, p-value= 0.169) show an inter-rater agreement really high.

5.Automatic Syntactic Simplification system

5.1 Analysis stage

This stage takes the original sentence as input and decides whether simplifying it or not. It mainly relies on Part of Speech (PoS) tagging, word searches and, in a specific case, lemmatization. PoS taggers are only necessary in order to find the number of verbs in the sentence. As explained before, this research will include the simplification of sentences containing two conjugated verbs, since sentences with only one conjugated verb will be considered as simples already. This stage was based on a decision-tree model taking into consideration indicators (Medero and Ostendorf, 2011) of possible simplification but also indicators suggesting not to simplify (*cuando, donde*…).

5.2 Transformation stage

Once the analysis stage has decided to simplify the input sentence, the transformation stage comes into place. At this stage, the system basically chooses a position to split the sentence and introduces new connectors if necessary. This differs a bit from the transformation stage of Siddharthan's text simplification architecture (Siddharthan, 2002), where the analysis stage were the one selecting the boundaries between clauses and the main sentence. This stage does not take into account the relation within the sentences nor the anaphoric references to be considered in the output sentences. This will be considered later on the regeneration stage.

5.3 Regeneration stage

This is the last and most challenging task of the SS process. After the splitting of the sentence on the transformation stage, many assumptions need to be taken in order to make these new simplified sentences readable and understandable. For instance, there are some anaphoric references which may be lost during the transformation process and they need to be fixed on this stage. For the anaphoric references and noun phrase repetition within the sentences, this stage relies mainly in noun chunking rather than dependency structures (Siddharthan, 2011). The sentence reordering is based on a simple algorithm regarding the original position of the sentences to be split. This algorithm is improved for the complicate case of non-restrictive relative clause containing embedded structures.

6. Evaluation

6.1 Syntactic Simplification for Machine Translation

In this Sub-section, the effect of SS in MT will be evaluated by a comparison between direct MT systems and MT systems which previously have used SS on the source languages. There will be two different evaluations concerning different target languages, the first one taking Korean as the target language and the second one taking Chinese as the target language. Spanish will be in both cases the source languages. The experiments settings are similar for both experiments. Thirty complex sentences have been randomly selected from thirty different texts of the parallel corpus created along this research (see Section 4), including their respective simplified ones. The thirty original sentences were translated respectively by a Spanish-Korean and a Spanish-Chinese translator, which was considered as Gold Standard. The Spanish-Korean and Spanish-Chinese Google Translate was the MT system for the evaluation. First, the original sentences were introduced directly on the automatic translator and the output sentences were stored, what we will call simply MT. Second, SS was used prior to the translation. This second output will be called SS+MT. The simplified part of the thirty original sentences (taken from the parallel corpus) was introduced on Google Translate to produce three new outputs. An evaluation of both ways was then carried out among three native Korean and three Chinese. The native speakers were asked to evaluate the grammaticality of the output sentence and compare each output with the Gold Standard regarding its preservation of meaning. The grammaticality was evaluated by following a 1-5 scale where 1 indicates a totally non grammatical sentence and 5 indicates a fully grammatical sentence. The preservation of meaning was also done by following the 1-5 scale in a similar way. 1 means that the output sentence does not preserve at all the meaning of the Gold Standard sentence and 5 indicates a total preservation of meaning.

There are some preliminary results obtained from three native Korean evaluators holding at least a university postgraduate degree. Spanish-Korean MT systems have not been as developed as other pair of languages. According to the native Korean evaluators of this research, the output

sentences given by the Spanish-Korean Google Translate MT system are poorly constructed and in many cases not understandable. This is reflected on the results which are summarized on Table 2.

Table 2: Spanish-Korean MT results

Spanish-Korean	**Grammaticality**		**Meaning preservation**	
	MT	**SS+MT**	**MT**	**SS+MT**
Average	2.6	3.07**	2.44	3.02**
Positive (4-5)	15.6%	31.1%	18.9%	28.9%
Neutral (3)	40%	44.4%	24.4%	35.6%
Negative (1-2)	44.4%	24.4%	56.7%	35.6%

**: difference statistically significant at the level 0,001

According to a paired t-test, the difference between the grammaticality score averages of MT and SS+MT output sentences is statistically significant ($t(89)=-5.205$, p-value<0.001). This is a considerable difference taking into account that the Gold Standard sentences have been translated directly from the original sentences and not the simplified ones. The results from the preservation of meaning task are also summarized on Table 3. In this case, the meaning is better preserved by SS+MT. The difference is also statistically significant according to a t-test ($t(89)=-6.183$, p-value<0.001).

As far as the Spanish-Chinese MT experiment is concerned, some preliminary results were obtained from three graduated native Chinese speakers. The results obtained from the evaluation about the grammaticality task are summarized on Table 3. According to the results, the grammaticality is improved when applying SS prior to MT. The difference is statistically significant at the 0.05 level according to a paired t-test ($t(89)=-2.074$, p-value=0,041). The meaning preservation results were even more promising than the one regarding the grammaticality task. The meaning preservation average obtained by applying direct MT got the

extremely low score 1.82, with 81.1% of the sentences obtaining a negative attitude from the evaluator. These results, even being still low for a MT system, show an important improvement by applying SS prior to the translation. The average reaches a 2.18 score and the percentage of sentences with a neutral or positive attitude by the evaluators is considerably raised, whereas the percentage of sentences with a negative attitude from the evaluator decreases until 65.6%. The average difference between MT and SS+MT for the meaning preservation task is even statistically more significant that in the grammaticality task according to an paired t-test ($t(89)=-3,734$, p-value<0.001).

Table 3: Spanish-Chinese MT results

Spanish-Chinese	**Grammaticality**		**Meaning preservation**	
	MT	**SS+MT**	**MT**	**SS+MT**
Average	2.41	2.62*	1.82	2.18**
Positive (4-5)	12.2%	12.2%	4.4%	8.9%
Neutral (3)	31.1%	44.4%	14.4%	25.6%
Negative (1-2)	56.7%	43.3%	81.1%	65.6%

*: difference statistically significant at the level 0,05

**: difference statistically significant at the level 0,001

6.2 Foreigners learning Spanish

Twenty four Spanish learners at the Faculty of *Traducción e Interpretación* in Granada University agreed to collaborate in this project. The participants had a native language other than Spanish. They were divided in two groups of twelve participants each: A and B. Group A participants were given two original texts from AnCora Corpus and were asked to answer eight multiple-choice questions about them. It was a reading-comprehension test. Group B

students were asked to do the same but from the group A's simplified texts. These simplified texts belong to the parallel corpus created along this research. The participants of both groups had a similar level of English: five participants with an advanced level (C1-C2 according to the Common European Framework of Reference for Languages) of Spanish and seven with an Intermediate level (European Framework of Reference for Languages).

The results obtained from the test were processed and summarized on Table 4. Surprisingly, the participants from group A (original texts) got a higher amount of right answers than the participants from group B (simplified texts). Group A obtained 82.3% of correct answers in contrast to the 79.2% obtained by group B. However, after applying an un-paired t-test at 0.05 level ($t(22)=-0.457$, p-value=0.652), we concluded that this difference is not statistically significant. Therefore these results suggest that the text comprehension has not been affected by SS.

Table 4: Results from the reading-comprehension test

Correct answers	A: Original text	B: Simplified text
Total	82,3%	79,2%
Intermediate	82,1%	74,5%
Advanced	82,5%	87,5%

6.3 Automatic Syntactic Simplification system

The test set selection (40 sentences containing two conjugated verbs) was done by following a random process. Every sentence was randomly selected from an AnCora Corpus text. There is no more than one sentence selected from a specific text. Therefore 40 different texts were necessary for the extraction of all the sentences to be used as a test set. Since these 40 texts belong to the original part of the parallel corpus, the simplified part of the corpus was taken as a reference (Gold Standard) in order to evaluate the

system. The evaluation was then divided in three parts, one for each different stage.

The analysis stage takes an individual sentence from the test set as input and decides whether simplifying the sentences or not. The output of a single sentence takes only two values: True for a sentence to be simplified and False for a sentence which does not need simplification. As we can appreciate from Table 5, there were 12 sentences which were not simplified by the system and 28 correctly identified as simplifiable. In terms of precision, the analysis stage obtains a remarkable 100%. This means that there is no a single sentence simplified when it should not be simplified.

Table 5: Analysis stage results

	Simplification	No Simplification
Gold Standard (n. of sentences)	30	10
Correct	28	10
Wrong	2	0
Precision: **100%**		
Recall: **93.3%**		

The splitting of the original sentence is carried out on the transformation stage. There are 28 sentences from the test set to be split at this stage. The system correctly selects the positions of the clauses and the simplification cases in 100% of the sentences from and makes just a single mistake in the transformation. Therefore, the **accuracy of the transformation stage reaches 96.43%**, which is quite promising at this point.

For the regeneration stage, the same test set is used. 27 sentences are already correctly split on the transformation phase and need to be fixed at this stage. A general evaluation was done for the regeneration stage. Each output from the 27 already split sentences was given a 0-2 score. 2 means that the output is exactly same as the Gold Standard. 1 means that the output has the same

meaning as the Gold Standard but something has been modified on a different way, just affecting the readability. 0 means that the meaning of the original sentence is on some way changed by the regeneration. The results obtained by the regeneration module are shown on Table 6.

Table 6: Regeneration stage results

Score	0	1	2
Number of sentences	1	8	18
Percentage	3.7%	29.6%	66.7%
Average: 1.63			

As we can appreciate on Table 6, 18 out of the 27 sentences (66.7%) were correctly regenerated according to the Gold Standard, which is a remarkably high percentage taking into account the complexity of this module. The score average of the system is 1.63. This result is quite promising, as a few improvements may lead to an almost perfect score (2) by the system. To sum up, the general results obtained by the system are shown on Table 7. The general results obtained are promising, as 92.9% of the simplified sentences were positively handled by the system.

Table 7: Simplification system general results

	Total	Percentage
Input sentences	40	100%
Sentences simplified	28	70%
Positively simplified	26	92.9%
Incorrectly simplified	2	7.1%
Perfectly simplified	18	64.3%

Sentences not simplified	12	30%

7. Conclusions and future work

The results from the evaluation state that a reliable syntactic simplification system is feasible by considering splitting rules proposed on this paper. A few adjustments need to be made especially on the regeneration stage, focusing on particular features from the Spanish language. Further research should focused on the development of the Syntactic Simplification system and extend it to handle every kind of sentence from the natural language. Different rules could be also added in the system and evaluate its impact on extrinsic applications other than Machine Translation (Text Summarization, Information Extraction and parsing, for instance). Modules at the lexical, discourse and semantic level could be implemented in the system in order to solve some ambiguity problems on these extrinsic applications. A possible way to improve the system will be to extend the parallel corpus in order to apply *Machine Learning* techniques. The future system could be hybrid by using the syntactic rules and Machine Learning in some complicated cases.

As far as Machine Translation is concerned, Syntactic Simplification has been proved to be a quite reliable and easy to implement monolingual resource for underdeveloped systems. Spanish-Chinese and Spanish-Korean Statistical Machine Translation systems have been evaluated on this research. The preliminary results state that the application of Syntactic Simplification prior to the translation has been proved to be really beneficial on both cases. However, other languages should be taken into consideration as target languages in the future. Further research should also focus on finding accurate evaluation metrics, either automatic or human, for the task.

Acknowledgements

This research was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

I would like to acknowledge my supervisors: Prof. Sandrine Fuentes from Universitat Autònoma de Barcelona; Prof. Richard Evans from the University of Wolverhampton; and Prof. Sylviane Cardey-Greenfield from

Université Franche-Comté. I would also like to thank to all the people who helped me in some way during my stay in the three universities.

Thank you to Obra Social "La Caixa", which partially supported this masters by a "La Caixa" grant for master studies in Spain.

References

Aluísio, S., Specia, L., Pardo, T., Maziero, E., and Fortes, R. (2008) Towards Brazilian Portuguese Automatic Text Simplification Systems. ACM *Symposium on Document Engineering*, pp. 240–248.

Bosque, I. (dir.) (2004) *Redes. Diccionario combinatorio del español contemporáneo*, Ediciones SM, Madrid.

Bott, S. and Saggion, H. (2011) Spanish Text Simplification: An Exploratory Study. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural.*

Bott, S. and Saggion, H. (2011) An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. *Proceedings of the ACL Workshop on Monolingual Text-to-Text Generation*, pp. 20-26.

Camacho Collados, J. (2013) Splitting complex sentences for Natural Language Processing applications: Building a Simplified Spanish, *V International Conference on Corpus Linguistics*, Alicante, Spain, in press.

Cardey, S., Greenfield, P., Alsharaf, H., and Shen, Y. (2004) Problems and Solutions in Machine Translation Involving Arabic, Chinese and French. *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04).*

Cardey, S. (2011) Machine Translation of Controlled Languages for More Reliable Human Communication in Safety Critical Applications. *Proceedings of the 12th International Symposium on Social Communication - Comunicación Social en el Siglo XXI*, Santiago de Cuba, Cuba, Vol. II, pp. 953-958.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996) Motivations and methods for text simplification. *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pp. 1041–1044.

Evans, R. (2011). Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, 26(4), pp.371-388.

Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013), Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. *Computational Linguistics and Intelligent Text Processing*, 7817, pp. 488-500.

Goldberg, M. (1999). An Unsupervised Model for Statistically Determining Coordinate Phrase Attachment. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. College Park, Maryland.

Kaji, H. (1999) Controlled languages for machine translation: state of the art. *Proceedings of MT Summit VII "MT in the Great Translation Era"*. Kent Ridge Digital Labs, Singapore, pp. 37-39.

Klebanov, B. B., Knight, K., and Marcu, D. (2004). Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems*. Berlin: Springer-Verlag, pp. 735–47.

Koehn, P. (2010) *Statistical Machine Translation*. Cambridge University Press. Textbook.

Medero, J. and Ostendorf, M. (2011). Identifying Targets for Syntactic Simplification. *ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2011).*

Mitamura, T. (1999) Controlled Language for Multilingual Machine Translation. *Proceedings of Machine Translation Summit VII*.

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

Siddharthan, A. (2002) An Architecture for a Text Simplification System. *Proceedings of the Language Engineering Conference (LEC'02)*: Hyderabad, India, December 13-December 15, IEEE Computer Society, London, United Kingdom, pp. 64.

Siddharthan, A. (2011). Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different Generation Strategies. *Proceedings of the 13th European Workshop on Natural Language Generation*. Nancy, France.

Specia, L., Caseli, H.M., Pereira, T.F., Pardo, T.A.S., Gasperin, C., and Aluísio, S.M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh (ed), *Advances in Computational Linguistics, Research in Computer Science*, 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009). Mexico City.

Specia, L. (2010). Translating from Complex to Simplified Sentences. *9th International Conference on Computational Processing of the Portuguese Language (Propor-2010)*. Lecture Notes in Artificial Intelligence, 6001, Springer, pp. 30-39. Porto Alegre, Brazil.

Taulé, M., Martí, M., and Recasens, M. (2008) AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Language Resources and Evaluation-LREC 2008.*

Temnikova, I. (2012) *Text Complexity and Text Simplification in the Crisis Management domain*. Ph. D. Thesis. University of Wolverhampton, United Kingdom.