SCAP-TT: Tagging and lemmatising Spanish tourism discourse, and beyond

Patrick Goethals, Els Lefever and Lieve Macken
Ghent University (Belgium)
patrick.goethals@ugent.be, Els.Lefever@ugent.be &
Lieve.Macken@ugent.be

Abstract

In this research note we report on the first results of *SCAP*, the *Spanish Corpus Annotation Project,* applied to tourism discourse. In particular, we present and assess a new *TreeTagger* parameter set for Spanish (*SCAP-TT*), which has been trained for the Part-of-Speech tagging (POS-tagging) and lemmatisation of Spanish promotional tourism texts. Although *SCAP-TT* has been trained for specialized tourism discourse, we also show promising results for the annotation of other text genres such as essays and literary texts.

Keywords: POS-tagging, lemmatisation, Spanish, *TreeTagger*, tourism discourse, *SCAP*.

Resumen

SCAP-TT: El etiquetado gramatical y la lematización del discurso turístico español, y más allá

En esta nota de investigación describimos los primeros resultados de *SCAP*, el *Spanish Corpus Annotation Project*, aplicado al discurso turístico. Presentaremos y evaluaremos una nueva versión para el español del etiquetador *TreeTagger* (*SCAP-TT*), diseñado específicamente para el etiquetado y la lematización de textos turísticos promocionales en español. Si bien *SCAP-TT* ha sido diseñado para el discurso especializado del turismo, mostraremos también resultados muy prometedores para la anotación de otros tipos de discursos, como ensayos y textos literarios.

Palabras clave: etiquetado gramatical, lematización, español, *TreeTagger*, discurso turístico, *SCAP-tur*.

1. Introduction

This research was motivated by two observations. The first of them is that in Spanish specialized discourse corpus compilation projects, POS- and lemma-annotation are not yet self-evident features. Corpora often consist of raw text, allowing for word form-based queries, but not for more abstract POS- or lemma-based queries. Regarding Spanish tourism discourse, for example, the two main corpus projects, *Linguaturismo* (http:// www.linguaturismo.it) and *Cometval* (http://www.uv.es/cometval) do not (yet) contain linguistic annotations. This observation is not intended as a criticism towards these specific projects, but rather as one example of a broader dichotomy between current practices in corpus and computational linguistics.

The second observation is related to *TreeTagger*[1] (*TT*, Schmid, 1994, 1995). *TT* is a tool for automatic POS-tagging and lemmatisation which predicts the most probable POS-tag for each word taking into account its inherent formal characteristics and the surrounding POS-context. *TT* can be run using the built-in parameters, but it also offers a training tool to generate new parameter sets, which means that it can be adapted and improved depending on the specific needs of a corpus project. Although the main architecture is language-independent, the output quality varies according to the language, since the tool depends on language-specific input, such as a lexicon, a tag set, a list of multi-word items or a training corpus (for technical details, see Schmid, 1994, 1995). It is generally accepted that the results for the Spanish *TreeTagger* are not as good as for English, for example (Göhring, 2009). Moreover, it should be noted that the adaptiveness of *TreeTagger* appears to be underused, at least for Spanish, since there are no newly trained and publicly available parameter sets for Spanish.

Taking into consideration these observations, our aim is to use the inherent adaptiveness of TreeTagger and to make an improved parameter set for Spanish. In order to stimulate the development of annotated corpora, the parameter set is made available at the project's website (www.scap.ugent.be). At the same website, readers will find further technical information, as well as advanced tools and automated applications for further processing the TT-output. In what follows, we will first briefly discuss the performance of the current Spanish *TreeTagger* parameter set (*Standard-TT*). Then, we will describe the main decisions that were taken in the development of a new parameter set (*SCAP-TT*), and compare the results of *SCAP-TT* with *Standard-TT*. Finally, it is important to emphasize that in this research note,

we will not compare the results of *TreeTagger* with those of other tagging tools, such as *IULA* (Martínez et al., 2010), *GRAMPAL* (Moreno & Goni, 1995) or *FREELING* (Carreras et al., 2004) (see e.g. Parra & Martínez, 2015 for a recent comparison).

2. *TreeTagger*

At least three features determine the success of a tagging tool among corpus linguists: its user-friendliness, the accuracy of the output and the granularity (or level of detail) of the coding categories. As has been argued by several authors (e.g. Durán, 2010; Argüelles & Muñoz, 2012; Crespo & Frías, 2015), *TreeTagger* scores high on user-friendliness. It is free and easily downloadable, it can be run locally, it can be used with a high number of languages and it does not require advanced technical skills, especially since the publication of graphic interfaces by Duibhín[2] and Anthony (*TagAnt*[3]). Although some of these arguments might seem trivial for computational linguists, they are essential for corpus linguists, and the fact that *TreeTagger* scores high on user-friendliness basically guarantees that it will continue to be used in the future.

However, with respect to accuracy, the results of the current Spanish *Standard-TT* are sub-optimal. Reports vary between 90% and 95% accuracy, depending on the text type and on how accuracy is measured. In a recent comparison of several tools (*TreeTagger*, *IULA*, *Freeling* and *IXA*), Parra and Martínez (2015) conclude that *Standard-TT* ends third out of four both for POS-tagging and lemmatisation. Also with respect to granularity, *TT* does not score optimally, mainly because it does not include inflectional information.

In this contribution, the focus is on improving the accuracy of both POS-tagging and lemmatisation. In a later stage, when more training data will be available, we will attempt to also improve the granularity of the coding scheme.

3. Design of the new parameter set

In what follows we describe the main steps taken to generate a new parameter set for Spanish. This information is not exhaustive: detailed and explicit coding guidelines can be found at the project's website (www.scap.ugent.be).

3.1. Tokenization and multi-word units

A first decision concerns an optional *TT*-feature that affects the definition of multi-word units. *TreeTagger* defines multi-word units (e.g. "sobre~todo") during the tokenisation process, on the basis of a closed list (approx. 300 items in Standard-TT). This procedure presents some inherent weaknesses. First, the built-in list seems relatively arbitrary, including for example "tal~vez", "a~veces", and "San~Pablo" but not "a~lo~mejor", "de~vez~en~cuando" or "Santa~María". The end user can partly overcome this problem by adding or omitting items. The second weakness, however, is more problematic and is caused by the fact that multi-words are defined before the contextual analysis takes place (i.e. during the tokenisation process). This means that all instances of these word sequences are a priori defined as multi-words. In a sentence such as "sobre todo el territorio", "sobre todo" would be analysed as a multi-word unit, just like in "es, sobre todo, muy inteligente". Therefore, we took the decision not to use a list of multi-word items. The end-user of *TreeTagger* is still free to use (and modify) such a list, but we do not recommend this, especially not with *SCAP-TT*, since it is not trained this way.

3.2. POS tag set: Accuracy and granularity

The total tag set increased from 77 tags in *Standard-TT* to 110 in *SCAP-TT*. Although the number of tags is sometimes used as an indicator of granularity, our main reason for remodelling some aspects of the tag set was to improve the accuracy. *SCAP-TT* scores at least as good, and perhaps slightly better, on granularity than *Standard-TT*, but it still lacks inflectional information.

In what follows, we discuss three decisions to illustrate the overall philosophy. As said before, a detailed comparison between the two tag sets can be found at the project's website (www.scap.ugent.be).

3.2.1. Past participles

The classification of adjectival and verbal uses of past participles is a well-known problem, which is handled differently by different taggers (Parra & Martínez, 2015). The distinction causes frequent errors (e.g. when "adecuada" in "una medida adecuada" is analysed as a verb and lemmatised as "adecuar"), and infelicitous decisions (e.g. when "pasado" receives the

same label in "los tiempos pasados" and in "ha pasado algo"). The consequence of this problematic labeling is that both categories of adjectives and verbal participles are contaminated and should be entirely corrected during post-edition, which is very time-consuming. As a pragmatic solution, Parra and Martínez (2015) decide to classify all past participles as verbal forms, and a similar option is taken in the *EAGLES* tagset.[4] In fact, the problem is also conceptual, since it seems artificial to divide what is a continuum ranging from prototypical adjectival uses to prototypical verbal uses in only two categories. Given both the pragmatic and conceptual arguments, we decided to create three categories:

(i) **ADJ**, for adjectives not derived from a participle;

(ii) **VPART**, for participles in their most prototypical verbal contexts, namely in combination with "haber" and in passive constructions with "ser";

(iii) **ADJV**, for adjectives derived from a participle that do not occur in combination with "haber" and passive "ser".

As a result, the categories ADJ and VPART are far more homogeneous, and the end user can choose (i) whether it is necessary to check the intermediate category of ADJV or (ii) to add the whole category of ADJV to ADJ or to VPART.

3.2.2. Enclitics

TreeTagger does not separate verbal forms and enclitics, but assigns a specific POS-tag to verbs containing one or several clitics. In *SCAP-TT*, this option is further refined, assigning different labels to the four possibilities:

(i) **SE** when the verb is followed by "se"("arrepentirse");

(ii) **CLI** when the verb is followed by one pronoun ("comprarlo");

(iii) **SECLI** when the verb is followed by "se" and a pronoun ("decírselo");

(iv) **CLI2** when the verb is followed by two pronouns ("comprármelo").

3.2.3. Articles

In contrast with *Standard-TT*, which has only one label for articles, *SCAP-TT* distinguishes between definite, indefinite and neutral ("lo más importante") articles.

3.3. Lexicon

SCAP-TT uses a lexicon of almost 79,000 lemmas and 670,000 word form – POS tag pairs. This lexicon combines various sources, but takes as a starting point the stemming list made available by Boleslav Měchura[5] (containing some 490,000 word form – lemma pairs). The latter list lacked important information: it did not include invariable forms (e.g. adverbs), word forms that coincide with the lemma, verbal forms with enclitics, and, most importantly, information on the POS-category of the word forms. In the *SCAP* lexicon all this information was added.[6]

3.4. Training corpus

The *SCAP* parameter set was trained on a manually annotated 200,000 word corpus containing two types of descriptive-promotional tourism texts: digital descriptions of tourism attractions published in *Minube*, an online 2.0 travel guide, and brochures published by *TurEspaña*, the national tourism agency.

4. Testing and results

SCAP-TT was tested on three 5,000 words-corpora. The first testing corpus belongs to the same series of tourism brochures as those used in the training corpus. There is, however, no textual overlap between testing and training data. Additionally, we considered two other 5,000 words excerpts from the essay *Las venas abiertas* (E. Galeano), and from the novel *El Club Dumas* (A. Pérez Reverte). The testing data allow us to assess the added value of *SCAP-TT*, both in the specialized context for which it is trained, and in other discourse domains.

Two preliminary observations should be made. First, it is important to note that we applied an assessment procedure and not a gold standard procedure, which means that we evaluate accuracy with respect to the internal logic of the parameter set, and that tags are only counted as errors when they do not correspond to what is expected from the tag set definitions. This avoids that

the error analysis gets biased by judgments about the felicitousness of specific coding decisions. For example, "nadie" is systematically tagged by *Standard-TT* as a Quantifier, while it is a Pronoun in *SCAP-TT*. Although we believe that the second label is more appropriate than the first one, we did not count this as an error in the *Standard-TT* output. On the other hand, when a participle following "haber" is tagged as VLadj in *Standard-TT* and as ADJV in *SCAP-TT*, this is considered correct in the former case, but erroneous in the latter one, because in *SCAP-TT* we would expect to find VLPART. Secondly, we will take into account the effect of using the optional TT-feature of "Capitalization heuristics". Using this heuristics means that the tagger seeks unknown capitalized words in the lower-case lemma list, which may affect the ratio of proper nouns in the tagging output, and also the type of errors that occur. Therefore, errors are subdivided in categories: (a) erroneous POS-tags excluding proper nouns, (b) items which receive a correct POS-tag but are not appropriately lemmatised, again excluding proper nouns, (c) proper nouns which were not recognized as such, and (d) items which were incorrectly labelled as proper nouns. Two totals are provided: one excluding the proper nouns, and one overall total.

The tourism testing corpus clearly shows that *SCAP-TT* improves POS-tagging and lemmatisation substantially, especially when proper nouns are excluded. The number of errors is even reduced with more than 80% (30 vs. 215 or 37 vs. 271). In addition, not using the Capitalization heuristics reduces the number of missed proper nouns drastically (27 vs. 135), although it also slightly increases the number of general POS-errors (27 vs. 17) and the number of false proper nouns (11 vs. 5). These are important results, but the question arises whether the improvements are only due to the specialized training modalities. Yet, the results for the other testing corpora show that *SCAP-TT* yields better results in literary texts and essays as well, although the benefit is less pronounced than in the tourism domain. For example, considering the <–Cap Heuristics> modus, the total number of accuracy fails is reduced from 7,2% to 1,5% in the tourism testing corpus, from 6,2% to 2,8% in the literary corpus and from 4,1% to 2,7% in the essay corpus. We also notice that the most significant improvements in the literary and essay corpus concern lemmatisation, and that the use of capitalization heuristics leads to considerable shifts in the results of *SCAP-TT*. Based on these results, the best strategy is to combine both outputs by replacing in the <+Cap. Heuristics> output those POS-labels in which the – capitalization heuristics tags a proper noun.

	Standard-TT		SCAP-TT	
	+ Cap. Heuristics	- Cap. Heuristics	+ Cap. Heuristics	- Cap. Heuristics
Test 1: Tourism Brochure				
(a) POS-error (no proper nouns)	133 (2.7%)	134 (2.7%)	17 (0.3%)	27 (0.5%)
(b) correct POS but erroneous or unknown lemma (no proper nouns)	82 (1.6%)	83 (1.7%)	13 (0.3%)	10 (0.2%)
(c) missed proper nouns	159 (3.2%)	141 (2.8%)	135 (2.7%)	27 (0.5%)
(d) false proper nouns	3 (0.1%)	3 (0.1%)	5 (0.1%)	11 (0.2%)
TOTAL (no proper nouns)	215 (4.3%)	271 (4.3%)	30 (0.6%)	37 (0.7%)
TOTAL	377 (7.5%)	361 (7.2%)	170 (3.4%)	75 (1.5%)
Test 2: Literary Prose				
(a) POS-error (no proper nouns)	132 (2.6%)	133 (2.7%)	58 (1.2%)	102 (2.0%)
(b) correct POS but erroneous or unknown lemma (no proper nouns)	107 (2.1%)	108 (2.2%)	5 (0.1%)	5 (0.1%)
(c) missed proper nouns	72 (1.4%)	68 (1.4%)	76 (1.5%)	29 (0.6%)
(d) false proper nouns	2 (0.0%)	2 (0.0%)	2 (0.0%)	6 (0.1%)
TOTAL (no proper nouns)	239 (4.8%)	241 (4.8%)	63 (1.3%)	107 (2.1%)
TOTAL	313 (6.3%)	311 (6.2%)	141 (2.8%)	142 (2.8%)
Test 3: Essay				
(a) POS-error (no proper nouns)	88 (1.8%)	91 (1.8%)	46 (0.9%)	54 (1.1%)
(b) correct POS but erroneous or unknown lemma (no proper nouns)	73 (1.5%)	74 (1.5%)	7 (0.1%)	7 (0.1%)
(c) missed proper nouns	68 (1.4%)	35 (0.7%)	107 (2.1%)	21 (0.4%)
(d) false proper nouns	1 (0.0%)	4 (0.1%)	5 (0.1%)	53 (1.1%)
TOTAL (no proper nouns)	161 (3.2%)	165 (3.3%)	53 (1.1%)	61 (1.2%)
TOTAL	230 (4.6%)	204 (4.1%)	165 (3.3%)	135 (2.7%)

Table 1. Testing results of *Standard-TT* and *SCAP-TT* for three corpora.

5. Conclusion

We have shown that *SCAP-TT* considerably improves the tagging and lemmatisation results of the current Spanish *TreeTagger*, especially but not

exclusively for tourism discourse. We believe that this is an important contribution since it may reinforce the use of an already well accessible and well-known tool and, as such, contribute to integrating POS-tagging and lemmatisation into the current practice of Spanish corpus researchers. Unsurprisingly, we have also found that the new tagger gives the best results for the specific discourse domain for which it is trained.

Acknowledgements

We wish to thank H. Schmid for the very helpful and quick answers to our practical questions regarding the use of the *TreeTagger* Training Tool. .

Article history:
Received 5 May 2016
Received in revised form 6 September 2016
Accepted 9 September 2016

References

Argüelles Álvarez, I. & A. Muñoz Muñoz (2012). "An insight into Twitter: A corpus based contrastive study in English and Spanish". *Revista de Lingüística y Lenguas Aplicadas* 7: 37-50.

Carreras, X., I. Chao, L. Padró & M. Padró (2004). "FreeLing: An open-source suite of language analyzers" in *Proceedings of _The Fourth International Conference on Language Resources and Evaluation, LREC* 2014, 239-242. Lisbon: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/271.pdf [xx/xx/xxxx]

Crespo, M. & A. Frías (2015). "Stylistic authorship comparison and attribution of Spanish news forum messages based on the TreeTagger POS tagger". *Procedia-Social and Behavioral Sciences* 212: 198-204.

Durán Muñoz, I. (2010). "A corpus-based ontoterminological tool for tourist translations". *International Journal of Translation* 22: 149-165.

Göhring, A. (2009). *Spanish Expansion of a Parallel Treebank*. MaThesis. University of Zurich.

Martínez, H., J. Vivaldi & M. Villegas (2010). "Text handling as a Web Service for the IULA processing pipeline". In *Proceedings of LREC 2010: Workshop on Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation*, 22-29. Paris: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W3.pdf [xx/xx/xxxx]

Moreno, A. & J.M. Goni (1995). "GRAMPAL: A morphological processor for Spanish implemented in PROLOG" in *arXiv preprint cmp-lg/9507004*.

Parra Escartín, C. & H. Martínez Alonso (2015). "Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task". *Procesamiento del Lenguaje Natural* 54: 29-36.

Schmid, H. (1994). "Probabilistic part-of-speech tagging using decision trees" in *Proceedings of The International Conference on New Methods in Language Processing*, 44-49.

Schmid, H. (1995). "Improvements in part-of-speech tagging with an application to German" in *Proceedings of The ACL SIGDAT-Workshop*, 1-9.

Patrick Goethals is Associate Professor at the Department of Translation, Interpreting and Communication, Ghent University (Belgium). His main research interests are Spanish linguistics, corpus-based translation studies, multilingual communication, and tourism communication. He has published several articles in international journals such as *Journal of Pragmatics*, *Meta*, *Linguistics*, *Ibérica* and *Multilingua*.

Els Lefever is Assistant Professor at the LT3 language and translation technology team at Ghent University. She has a strong expertise in machine learning of natural language and multilingual natural language processing, with a special interest for computational semantics, cross-lingual word sense disambiguation and multilingual terminology extraction. She teaches Terminology and Translation Technology, Language Technology and Digital Humanities courses.

Lieve Macken is Assistant Professor at Ghent University with strong expertise in multilingual language processing. Research interests are computer-assisted translation, terminology extraction, human-computer interaction in translation and machine translation. She is the operational head of the language technology section of the department, where she also teaches Translation Technology, Machine Translation, Localisation and Technical Translation.

NOTES

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[2] http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm.

[3] http://www.laurenceanthony.net/software.html.

[4] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html.

[5] http://www.lexiconista.com/datasets/lemmatization/.