BUILDING A SPANISH LEXICON FOR CORPUS ANALYSIS

RICARDO-MARÍA JIMÉNEZ-YÁÑEZ

Universitat Internacional de Catalunya, Barcelona

HUGO SANJURJO-GONZÁLEZ

Universidad de León, León

PAUL RAYSON

Lancaster University, Lancaster

SCOTT PIAO

Lancaster University, Lancaster

RESUMEN

El propósito de esta investigación es describir el proceso de creación de un lexicón en español anotado semánticamente que sirva para analizar corpus más amplios en lengua española. Los recursos semánticos más utilizados en la actualidad son WordNet, FrameNet, PDEV o USAS, pero se emplean principalmente para investigaciones relacionadas con la lengua inglesa. La creación de un lexicón semántico en español a gran escala posibilitará un aumento del tipo de estudios realizados a través del análisis de corpus en español. En la descripción de los pasos seguidos para la construcción del lexicón se muestran las distintas dificultades encontradas y las soluciones utilizadas para superarlas. Finalmente, la construcción del lexicón permitirá que investigaciones específicas como el análisis de metáforas, el análisis crítico del discurso o incluso disciplinas más alejadas como el procesamiento de lenguajes naturales, se beneficien notablemente.

Palabras clave: lexicón, Español, anotación semántica, Análisis del Discurso.

ABSTRACT

This paper seeks to describe the creation of a Spanish lexicon with semantic annotation in order to analyse more extensive corpora in the

Spanish language. The semantic resources most employed nowadays are WordNet, FrameNet, PDEV and USAS, but they have been used mainly for English language research. The creation of a large Spanish lexicon will permit a greater amount of studies of corpora in Spanish can be undertaken. In the description of the steps followed for the construction of the lexicon, the difficulties encountered in its creation, and the solutions used to overcome them will be described. Finally, the construction of the lexicon will allow specific research tasks to be carried out, such as metaphor analysis, ACD studies and even PLN studies.

Keywords: lexicon, Spanish, Semantic tagging, Discourse Analysis.

1. SEMANTIC TAGGING: BACKGROUND

The last two decades have seen the development of various semantic lexical resources such as WordNet (Princeton University, 2010), FrameNet (Baker, Fillmore and Lowe, 1998), PDEV (Hanks, 2014) and the USAS semantic lexicon (Rayson et al., 2004), which have played an important role in the areas of natural language processing and corpus-based studies. Semantic tagging has various applications in research areas such as metaphor analysis (Koller et al., 2008) and critical discourse analysis (Prentice, 2010) (Breeze, 2015) (Breeze 2016), however most of the research has focused on English language. Recently, efforts have been made to develop multilingual semantic lexicons, aiming to support multilingual/cross lingual corpus linguistics and natural language processing, and the development of a large-scale Spanish semantic lexicon will facilitate deeper corpus-based studies on Spanish language, and it will promote research in a wider range of areas such as corpus-based Spanish natural language processing.

2. CREATING AND EDITING INITIAL SPANISH LEXICON

In this paper, we report on the construction of a Spanish semantic lexicon, which employs the unified Lancaster semantic taxonomy and provides a lexical knowledge base for the automatic UCREL semantic annotation system (USAS). According to (Piao et al., 2016; Piao et al., 2015:1272), if appropriate high-quality bilingual lexicons are available, it is feasible to rapidly generate prototype semantic lexicons for a given language with a good lexical coverage. Following their approach, we have been constructing Spanish semantic lexicons targeting the generation of a high quality and large scale resource.

It is a challenging and time-consuming task to build semantic lexicons for new languages. In the beginning, the Spanish lexicon only contained 2,005 Spanish single-word entries automatically generated by translating the USAS English semantic lexicon entries using a Spanish-English dictionary of the top 5,000 words in Spanish compiled by Mark Davies (Davies, 2006). As a consequence of the automatic process, the Spanish lexicon contained some inaccuracies and errors. Therefore, a post-editing process was carried out to correct the entries in the lexicon. This process was made manually by a linguist, so he was able to review the applicability of the USAS taxonomy for Spanish.

2. 1 Post-Editing PROCESS

It is a very labour intensive exercise to manually check if the English lexicon entries are successfully transported to Spanish equivalent, because several issues had to be addressed.

It was found that most of the errors were related to English polysemy. An English word does not always correspond to unique Spanish word. For instance, ‘mine’, a possessive word in English, generally means ‘mío’ in Spanish, but in English ‘mine’ can also mean explosive, which is not linked to the word ‘mio’ in Spanish.

In addition, the Spanish POS tagset used in the bilingual lexicon is more complex than the simplified tagset employed in automatic translation process. For instance, in the simplified POS tagset there are no subcategories of pronoun or determiner, causing some meaning knowledge to be lost.

Lastly, many Spanish lemmas are not correctly recognised by the TreeTagger Part-Of-Speech Tagger (Schmidt, 1995) used in our work. Some words have multiple entries because some already tagged words are added as new words due to its different lemmas. For instance, the TreeTagger does not identify correct lemmas for those Spanish adverbs ending with "–mente", instead it puts their adjective forms as lemmas, which are not correct. A more detailed example is the Spanish word "constantemente", the TreeTagger tags it as an adverb (ADV) and puts as its lemma "probable" instead of "probablemente". We are considering replacing the TreeTagger in our framework with another tagger which does not suffer from these problems.

Word	**POS**	**Lemma**
La	ART	el
Policía	NC	policía
Recurred	VLfin	recurrir
A	PREP	a
Él	PPX	él
Constantemente	ADV	*constante*

Figure 1: Example of incorrect lemmatisation

2.2. INCREASING LEXICAL COVERAGE

In order to increase lexical coverage of the Spanish lexicon, some part-of-speech tagged corpus resources were used. Two main data sources include:

a) 1,000 most frequent words in Spanish language according to the information provided by CORDE corpus (RAE, 2016),

b) 660 most frequent words in religious scope from 4 editorial corpus of Spanish newspapers selected from PhD dissertation of Jiménez-Yáñez (2017).

The following process was carried out for incorporating these new words to the Spanish lexicon:

- A complete word list was extracted and analysed grammatically using TreeTagger software (Schmidt, 1995).
- Filter out some words as a result of grammatical errors as previously discussed.
- A unique lemma and the corresponding grammatical tag for each word was obtained.
- List of lemmas together with the grammatical tags were semantically tagged manually according to USAS

semantic category taxonomy
(http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf).

In addition, the words from the religious field were manually matched and compared with keywords from *Diccionario ideológico de la lengua española* (Casares, 1989, xxxvi-xxxvii). This process provided more accurate results in religious context.

According to (Piao et al., 2015), Spanish lexicon reached an average rate of 56.77% (Piao et al., 2015). After the expansion, the lexical coverage should be higher.

Development stages of semantic lexicon for Spanish are summarised in the following figure (Figure 2)

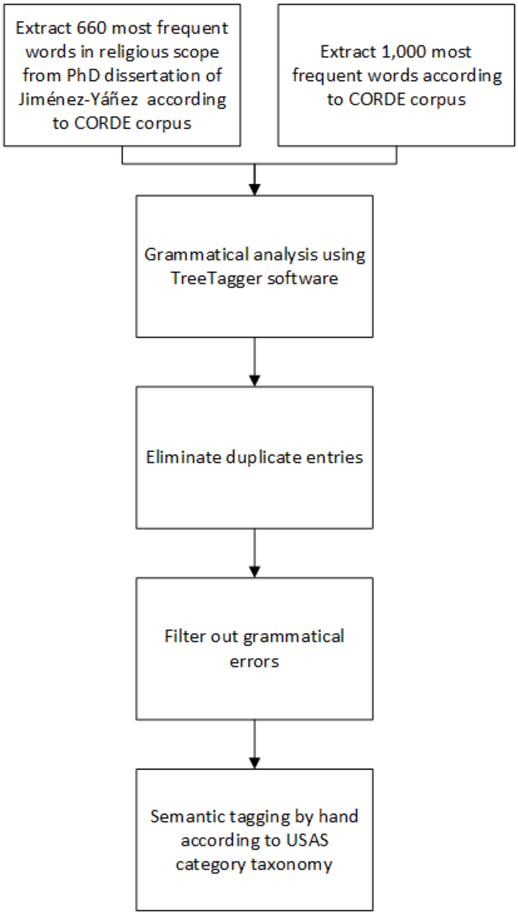
Extract 660 most frequent words in religious scope from PhD dissertation of Jiménez-Yáñez according to CORDE corpus
Extract 1,000 most frequent words according to CORDE corpus
Grammatical analysis using TreeTagger software
Eliminate duplicate entries
Filter out grammatical errors
Semantic tagging by hand according to USAS category taxonomy

Figure 2: Development stages of semantic lexicon for Spanish language

3. SPANISH SEMANTIC LEXICON AT PRESENT

After several rounds of expansion, such as the expansion in the business domain (Sanjurjo-González et al., forthcoming), current Spanish lexicon contains 4,206 words and 114 multiword expressions, as shown in Table 1.

Lexicon	**Single word entries**	**Multiword expressions**
Initial Spanish Lexicon	2,005	0
Current Spanish Lexicon	4,206	114

Table 1: Semantic lexicon sizes for Spanish.

We are aware that the Spanish lexicons need further expansion in order to achieve a high lexical coverage and need to be more precise in terms of semantic classification of the lexicon entries both on general and specific text types. Nonetheless, the current Spanish semantic lexicon already provides a useful resource for corpus-based studies on Spanish language.

The lexicon is available for academic use from website: https://github.com/UCREL/Multilingual-USAS.

Appendix provides a sample output of USAS Spanish tagger.

4. CONCLUSIONS

A fully developed Spanish semantic lexicon can be applied in various studies related to applied linguistics, such as the analysis and categorisation of the religious stance of an election manifesto, exploration of the semantic patterns of manifestos of populist parties, and more generally sentiment analysis or information about the style of writing.

APPENDIX

Example of USAS Spanish tagger output. USAS semantic tagset information is available in:
http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

TOKEN	LEMMA	POSTAG	SEMTAG
La	el	art_ART	Z5
semana	semana	noun_NC	N6+ T1.1 T1.3
santa	santo	adj_ADJ	S2 S9
es	ser	verb_VSfin	A3+ L1 Z5
la	el	art_ART	Z5
conmemoración	conmemoración	noun_NC	Z99
anual	anual	adj_ADJ	N6
cristiana	cristiano	adj_ADJ	S9 S9/S2mf Z1mf
de	de	prep_PREP	Z5
la	el	art_ART	Z5
Pasión	pasión	noun_NC	Z99
,	,	punc_CM	PUNCT
Muerte	muerte	noun_NC	L1- A5.1- E4.1-
y	Y	conj_CC	Z5 A1.8+
Resurrección	resurrección	noun_NC	B3 T2+/N6+ S9/A2.1+
de	de	prep_PREP	Z5
Jesús	Jesús	pnoun_NP	Z99
de	de	prep_PREP	Z5
Nazaret	Nazaret	pnoun_NP	Z99
.	.	punc_FS	PUNCT

REFERENCES

Baker, C. F., Fillmore, C. J., & Lowe, J. B. 1998, August. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and*

17th International Conference on Computational Linguistics 1: 86-90. Association for Computational Linguistics.

Breeze, R. 2015. “Ideology in corporate language: discourse analysis using Wmatrix3”. *CILC 2015 Seventh International Conference on Corpus Linguistics*. University of Valladolid, 5 – 7 March 2015.

Breeze, R. 2016. “Exploring the potential of semantic tagging for discourse analysis”. *CADS Conference*, Siena, 30 June – 2nd July 2016.

Casares, J. 1989. *Diccionario ideológico de la lengua española*. Barcelona, editorial Gustavo Gili, 2.ª edición (16.ª tirada).

Davies, M. 2002. «Un corpus anotado de 100.000.000 palabras del español histórico y modern». In: *SEPLN 2002 (Sociedad Española para el Procesamiento del Lenguaje Natural).* (Valladolid), 21-27.

Hanks, P. 2014. Pattern Dictionary of English Verbs (PDEV) – Project Page http://pdev.org.uk

Jiménez-Yáñez, R.-M. 2017. *La representación de la religión en editoriales de cuatro periódicos españoles (2009-2010).* Tesis inédita. Pamplona, Universidad de Navarra.

Koller, V., Hardie, A., Rayson, P. and Semino, E. 2008. “Using a semantic annotation tool for the analysis of metaphor in discourse”. *Metaphorik.de* 15: http://www.metaphorik.de/15/

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A. and Rayson, P. 2015. Development of the multilingual semantic annotation system. In proceedings of the *2015 Conference of the North American Chapter of the Association for Computational Linguistics* -

Human Language Technologies (NAACL HLT 2015), Denver, Colorado, United States, pp. 1268-1274.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez-Yáñez, R.-M., Knight, D., Křen, M. Löfberg, L. Adeel Nawab, R. M., Shafi, J., Lee Teh, P. and Mudraya, O. 2016. "Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages". In: *LREC 2016, Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), pp. 2614-2619. http://eprints.lancs.ac.uk/77965/

Prentice, S. 2010. "Using automated semantic tagging in Critical Discourse Analysis: A case study on Scottish independence from a Scottish nationalist perspective". *Discourse & Society*, 21(4): 405-437.

Princeton University. 2010. "About WordNet", WordNet. Princeton University. http://wordnet.princeton.edu

Real Academia Española. 2016. *Banco de datos (CORDE)* [en línea]. Corpus diacrónico del español. http://www.rae.es [19/09/2016].

Rayson, P., Archer, D., Piao, S. and McEnery, T. 2004. "The UCREL semantic analysis system". In Proceedings of *LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks*, pp. 7-12. Lisbon, Portugal.

Sanjurjo-González, H. *et alii*. forthcoming. "Extension of Spanish semantic lexicon in USAS: Issues and challenges".

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.