Multilingual CPA: Linking Verb Patterns across Languages

Vít Baisa[1], Sara Može[2], Irene Renau[3]
[1]Masaryk University; [2]University of Wolverhampton;
[3]Pontificia Universidad Católica de Valparaíso
e-mail: xbaisa@fi.muni.cz, S.Moze@wlv.ac.uk, irene.renau@pucv.cl

Abstract

This paper presents the results of a pilot study in linking corresponding English and Spanish verb patterns using both automatic and manual procedures. Our work is rooted in Corpus Pattern Analysis (CPA) (Hanks 2004, 2013), a corpus-driven technique that was used in the creation of existing monolingual pattern dictionaries of English and Spanish verbs, which were used in our experiment to design a gold standard of manually annotated verb pattern pairs. Research in CPA has inspired parallel projects in English, Spanish, Italian and German. Our study represents the first attempt to build a multilingual lexical resource by linking verb patterns in these languages. Verb have special difficulties related to grammar and argument structure that we do not find in other parts-of-speech, and for that reason we think that it is necessary to create a specific resource for them. After applying the automatic matching to a set of 87 Spanish verbs linked to 176 English verbs, an evaluation of a random selection of 50 of these pairs show 80% precision.

Keywords: Corpus Pattern Analysis; corpus lexicography; multilingual resources; verb patterns

1 Introduction

In this paper, we present the first steps in the creation of a multilingual lexical resource by linking verb patterns in existing pattern dictionaries using both automatic and manual procedures. The compilation of large, freely available multilingual lexical resources by means of linking pre-existing data has been gaining considerable traction in recent years, although much is yet to be done in order to improve such resources in terms of quality. Global WordNet (Vossen, 2002),[1] BabelNet (Navigli and Ponzetto, 2012),[2] Omega Wiki,[3] and Wiktionary[4] represent a step in the right direction in that they use word senses rather than words (or lemmas) to interlink the vocabulary of a number of different languages. As useful as these resources might be for the lexical description of nouns, none of them have successfully tackled the complexities of verb behaviour. Our ultimate goal is to fill this gap by providing a multilingual, corpus-driven lexical resource for verbs that can be used by language learners, language professionals (translators, editors) and the research community. We selected Corpus Pattern Analysis (Hanks 2004, 2013) as the methodological basis because it provides a technique to match verb meanings with syntagmatic contexts, using corpus evidence as the starting point.

The paper is structured as follows: first, the theoretical and methodological background underpinning

[1] http://globalwordnet.org/

[2] http://babelnet.org/

[3] http://www.omegawiki.org/

[4] https://www.wiktionary.org/

the research is presented (Section 2), followed by a detailed description of the linking technique developed in our pilot study (Section 3). This includes both the manual pattern linking procedure (Section 3.1) designed as the first step towards building a gold standard and the automatic linking algorithm (Section 3.2), which is evaluated in Section 4. Finally, our findings are summed up in the Conclusion and explore possible directions for our future work.

2 Background

Corpus Pattern Analysis (CPA) (Hanks 2004, 2013) is a technique in corpus linguistics and lexicography that associates word meaning with word use by mapping meaning onto specific syntagmatic patterns exhibited by a verb in any type of text. Based on the Theory of Norms and Exploitations (TNE), developed and presented by Hanks in a large number of publications (Hanks 2004; 2013; among others), CPA aims at identifying patterns of normal usage (norms), including literal and metaphorical uses, phrasal verbs and idioms, and exploring the way patterns are creatively exploited (exploitations) by means of painstaking lexical analysis of samples of corpus data. Its biggest contribution is in its effort to reflect real language use rather than preconceived speculations about language. In that respect, it provides a window into the normal, every-day phraseology, an area of study that has long been overlooked by fellow linguists who regarded patterns of normal usage as self-evident and therefore not worth scientific exploration. The technique is being implemented in the *Pattern Dictionary of English Verbs* (PDEV) (Hanks, in progress),[5] an online lexical resource that currently covers over 1,700 English verbs. Pattern dictionaries for languages other than English (Spanish, Italian,[6] German) are currently being compiled by fellow researchers across the globe. The *Pattern Dictionary of Spanish Verbs* (PSDV) (Renau, in progress) is an ongoing project offering around 100 high frequency Spanish verbs available online and constantly increasing this number.[7]
In CPA-based pattern dictionaries, verb entries are presented as lists of carefully described patterns, i.e. combinations of specific syntactic structures, lexical sets and semantic types representing typical syntactic role fillers for each pattern. For each verb, a random corpus sample consisting of at least 250 concordance lines is extracted from the British National Corpus[8] (Leech, 1992) and the Spanish Web Corpus[9] (Spanish), and tagged with pattern numbers using Sketch Engine (Kilgarriff et al., 2014). Larger samples (i.e. 500, 1000 or more lines) are used when dealing with particularly complex and/or frequent verbs. Patterns are identified mainly through lexical analysis of corpus lines, complemented by the information found in the Word Sketches i.e. lists of statistically relevant collocators and syntactic structures that are automatically generated using the functionality in the Sketch Engine. Patterns are then recorded and described in the CPA Editor (Baisa et al., 2015), our in-house lexicographic tool, using CPA's shallow ontology of semantic types (Ježek and Hanks, 2010), [10] which all CPA projects share.[11] Implicatures (pattern definitions) are written; register,

[5] http://www.pdev.org.uk

[6] The *Pattern Dictionary of Italian Verbs* is being developed by Elisabetta Ježek's team at the University of Pavia, Italy.

[7] PDSV is currently being compiled at the Pontifical Catholic University of Valparaíso, Chile. It is available online: http://www.verbario.com.

[8] http://www.natcorp.ox.ac.uk/

[9] https://www.sketchengine.co.uk/

[10] http://www.pdev.org.uk/#onto

[11] Nazar & Renau (2015) demonstrated that the CPA ontology can be successfully applied to the automatic population of a taxonomy of Spanish nouns.

domain, and idiom/phrasal verb labels are added, and links to FrameNet (Ruppenhofer et al., 2006)[12] are created, successfully linking the two complementary lexical resources. Dictionary entries also include quantitative information: for each separate pattern, a percentage is calculated based on the pattern's frequency in the annotated data.

Consider the verb *aggravate* shown in Figure 1. According to PDEV, this verb exhibits two patterns, which correspond to two separate verb meanings – the first relating to the deterioration of a state of affairs and the second to a person being annoyed by an event, state, process, or another person. The first pattern is by far the most common in every-day English, occurring in nearly 99% of the 235 concordances in the random sample.[13] In both patterns, the subject and the direct object are tagged with one or more semantic types from the CPA ontology: for instance, in the first pattern, only nouns corresponding to the semantic types [[Human]] or [[Eventuality]] (i.e. activities, processes, states, etc.) can occur in the subject slot. Implicatures (shown in blue) are anchored paraphrases of the pattern, describing the conventional meaning of the verb when used in a specific syntacto-semantic structure.

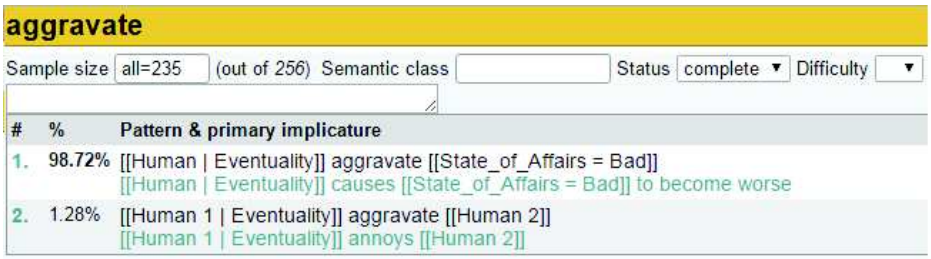

Figure 1: The dictionary entry for *aggravate* in PDEV, as shown in the CPA Editor.

Monolingual CPA-based dictionaries are highly compatible in that they are being compiled using the same tools and methodology. This presents researchers with a unique opportunity to create a multilingual lexical resource by means of linking corresponding patterns of verb use in two or more dictionaries using syntactic and semantic similarity as the deciding criteria. An additional advantage of cross-linguistic pattern linking is that monolingual pattern dictionaries are produced independently of each other, which prevents dictionary data from being skewed due to possible interferences between languages. If successful, this newly developed linking technique could make a significant contribution to the development of a whole new generation of multilingual lexical resources.

3 Methodology

3.1 Manual pattern linking

In our pilot study, we have conducted a manual linking task for a sample of English and Spanish verb patterns with the aim of identifying potential issues we will be facing in the future. We selected 87 Spanish verbs with one or more English equivalents (this resulted in a total of 126 English verbs): we included verb pairs such as *acusar-accuse* and semantically equivalent groups of near synonyms such as *enfadar-annoy/anger/infuriate/enrage.* The linked patterns were later used as a gold standard

[12] https://framenet.icsi.berkeley.edu

[13] As shown in Figure 1, the initial number of concordance lines in the extracted sample was 256 – the remaining 11 lines were discarded mostly due to errors in POS tagging and/or lemmatization.

for our automatic linking method. Only verbs exhibiting up to 15 patterns were included in the task in order to avoid the complexity of highly polysemous verbs, which deserve to be treated separately.
The linking task shed lights onto two important issues, the first being methodological and the second cross-linguistic. Both the English and the Spanish databases were created using the same methodology and tools, but unfortunately, they are still unfinished products that do not cover the same semantic groups of verbs, as researchers working on the two projects did not discuss a possible priority list of verbs that ought to be included in both databases. As a result, the number of potential matches is somewhat limited. Furthermore, our definition of a 'match' does not necessarily include close synonyms; for instance, *golpear* ('to hit') and *to stab* are listed as translation equivalents in some billingual dictionaries, but from a CPA-driven lexicogrammatical point of view, their semantic overlap is too low for us to consider them as potential matches.
Nevertheless, the thorniest issues we have encountered so far are related to the innumerable semantic differences between both languages – in other words, most candidate pairs do not exhibit full semanto-syntactic equivalence (also known as 'anisomorphism of languages' , cf. Yong and Pen (2007:135-173). Table 1 shows different types, or levels, of equivalence, from the exact match to different grades of equivalency connected to anisomorphism:

	Spanish pattern	**English pattern**	**Type of equivalence**
1	[[Human 1]] admirar [[Anything]]	[[Human]] admire [[Anything]]	*Exact match.* SP and EN ptts. share the same semantic types in the argument structure of the verb.
2	[[Light_Source | **Artifact**]] iluminar [[Physical_Object]]	[[Light_Source]] illuminate [[Physical_Object]]	*Partial match: additional sematic type.* Meaning-wise, both ptts. are very close, but one ptt. features an additional semantic type, i.e. [[Artifact]], which renders the SP ptt. semantically broader compared to the EN ptt.
3	[[Human | Eventuality]] estropear [[**Activity** | **Plan**]]	[[Eventuality 1 | Human]] spoil [[**Eventuality 2**]]	*Partial match: sematic types of different levels.* Meaning-wise, both ptts. are very close, with one of the corresponding semantic types being more general and therefore superordinated to the other semantic type in CPA's shallow ontology. In other words, [[Activity]] and [[Plan]] are types of [[Eventuality]]. This means that the SP direct object is semantically more specific than the EN counterpart.
4	[[Physical_Object]] aplastarse [NO OBJ]	([[Physical_Object]] crush [[Human]])	*Partial match: different alternations.* The two verb pairs differ in terms of the syntactic alternations they exhibit. Whereas the SP verb exhibits both inchoative and causative uses, the EN verb can only be used causatively. As a result, the inchoative SP ptt. cannot be linked with any ptt. of the EN target verb.

5	[[Anything]] deteriorar [[Character_Trait]]	-	*No match.* The SP ptt. cannot be matched to any ptt. exhibited by the EN verb *spoil* due to differences in meaning and syntax.

Table 1: Examples of candidate Spanish-English pattern pairs identified in the manual annotation task. Differences between pattern pairs are marked in bold. (SP ptt. = Spanish pattern; EN ptt. = English pattern).

In the manual annotation task, we considered cases similar to the examples in 1, 2 or 3 as matches, whereas cases such as 4 and 5 were marked as NPM ('no possible match'). The methodological and linguistic issues mentioned above prevented us from establishing full matches in a number of the cases we studied. Due to the complexity of the task, we decided to draft a simple decision tree that could help lexicographers decide whether or not a pair of patterns in two different languages can be considered a match. Figure 2 shows a preliminary version of the proposed decision tree.

3.2 Automatic linking

In order to improve the manual annotation task in terms of speed and precision, we implemented a heuristic-based algorithm to generate automatic linking suggestions. The dataset from the previous section was used as a basis in a preliminary evaluation of the proposed method.

To our knowledge, no similar automatic procedure for linking valency structures in two or more languages has been designed yet, except for a preliminary study on linking PDEV patterns to verb frames in the Czech valency lexicon VerbaLex (Hlaváčková and Horák 2005). Vonšovský (2016) proposes a similar approach to ours, splitting the task into a) matching ontology entries (semantic types and WordNet synsets) and b) comparing pattern (frame) structures.

3.2.1 Algorithm

For each of the 490 Spanish patterns, we computed a similarity score for all its possible translation into English (i.e. verbs and their respective patterns; a total of 5,067 Spanish-English pattern pairs). All candidate English patterns were sorted by the score and the top candidate, if available, was presented as output.

3.2.2 Similarity score

The similarity score was computed by comparing pattern structures in the two languages. Since this is preliminary work, our analysis only focused on the three main syntactic arguments, i.e. subject, direct object, and indirect object. It is important to note that a syntactic argument can be realized by nouns corresponding to more than one semantic type – for instance [[Human]] and [[Institution]] regularly alternate in the subject slot and are therefore often listed together. Whenever there was a non-empty intersection of semantic types in a given argument of the two patterns, each matched semantic type received one score point (only [[Human]], the most frequent semantic type, was assigned 0.5). If both given arguments were empty (also a match, mainly in the case of intransitive verbs), 0.5 score points were assigned. When the arguments contained different semantic types, the algorithm used the CPA ontology to check if the two types are in a hypernym relation (e.g. [[Event]] is the hyponym of [[Eventuality]]). The score for each hypo- or hypernym was based on their distance in the CPA ontology tree (the further apart they are located, the fewer score points they gain, measured in powers of 0.5). The three similarity scores (subject, direct object, indirect object) were

summed and the final score was assigned to the given pattern pair (cf. Table 2). Candidate pairs were sorted by the score and the top ranking pattern was returned.

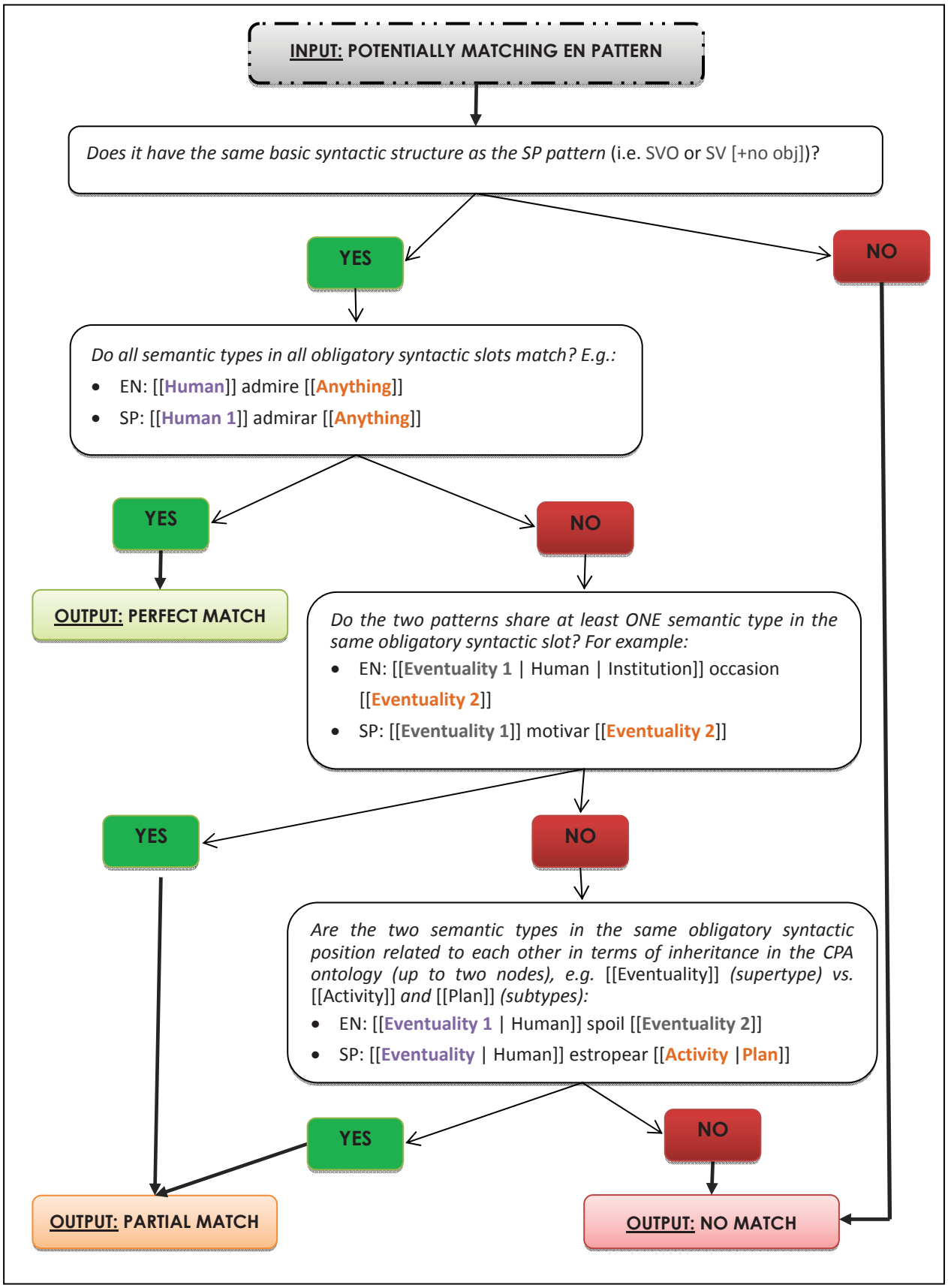

Figure 2: Preliminary version of the decision tree.

Slot Esp	Slot Eng	Score	Comment
[[Entity | Eventuality]]	[[Human]]	0.125	Human < Animate < Physical_Object < Entity, distance = 4
[[Human]]	[[Human]]	0.5	Human is almost in all patterns so the score was only 0.5
[[Artifact]]	[[Eventuality]]	0.0	No connection in the CPA ontology

Table 2: Ontology-based matching and the resulting scores.

4 Evaluation

To evaluate the procedure, we created a random sample of 50 Spanish-English verb pairs. We excluded all cases in which a Spanish pattern could not be matched against an English pattern in the sample, although we are fully aware of the fact that a matching English pattern could potentially be found outside the sample (we calculated that this happens in around 40% of the cases in our sample). Despite our work being at an early preliminary stage, the proposed method shows promising results, achieving 80% precision, as demonstrated in table 3.

	n	%
Correct matches	40	80
Incorrect matches	10	20
Total	50	100

Table 3: Results of the evaluation.

5 Conclusions and future work

The creation of a multilingual, CPA-based lexical resource is predicted to be a long-term, painstakingly slow and labour-intensive process that will involve partners from a number of institutions across the globe. In this paper, we presented the results of a preliminary study on linking patterns exhibited by Spanish verbs and their English counterparts using both manual and automatic procedures. In the future, we would like to develop larger bilingual lexicon fragments and use this newly annotated data as a gold standard dataset to train a robust automatic linking system. Our in-house dictionary writing tool, i.e. the CPA Editor, will also have to be adapted so that lexicographers will be able to effectively add cross-linguistic links between patterns.

6 References

Baisa, V., El Maarouf, I., Rychlý, P. & Rambousek, A. (2015). Software and data for Corpus Pattern Analysis. In Horák, A., Rychlý, P., and Rambousek, A. (eds.), *Ninth Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno. Tribun EU. 75–86.

Hanks, P. (2004). Corpus Pattern Analysis. In G. Williams & S. Vessier (Eds.), *11th Euralex International Congress. Proceedings.* Lorient: Université de Bretagne-Sud, pp. 87-97.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations.* Cambridge, MA: MIT Press.

Ježek, E., & Hanks, P. (2010) What lexical sets tell us about conceptual categories. *Lexis: E-journal in English lexicology. 4: Corpus Linguistics and the Lexicon*. Université Lumiere, Lyon. 7-22

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1): 7-36.

Hlaváčková, D., Horák, A. (2005). Verbalex–new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*.

Leech, G. (1992) 100 million words of English: the British National Corpus (BNC). *Language Research* 28(1):1–13.

Navigli, R. & Ponzetto, S. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence* 193: 217-250.

Nazar, R. & Renau, I. (2015). Ontology Population Using Corpus Statistics. In O. Papini, S. Benferhat, L. Garcia et al. (Eds.), *Proceedings of the Joint Ontology Workshops 2015 co-located with the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015).* Buenos Aires, Argentina, July 25-27, 2015.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R. & Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice. Berkeley, CA: ICSI.

Vossen, P. (2002). WordNet, EuroWordNet and Global WordNet. *Revue Française de Linguistique Appliquée* 7(1): 27–38.

Yong, H. & Peng, J. (1997). *Bilingual lexicography from a communicative perspective.* Amsterdam: John Benjamins.

Vonšovský, J. (2016). *Automatic Linking of the Valency Lexicons PDEV and VerbaLex* (master's thesis). URL: http://is.muni.cz/th/359500/fi_m/AutomaticLinking.pdf

Acknowledgements

This work has been partly supported by the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic, AHRC grant [DVC, AH/J005940/1, 2012-2015], project Contract nr. MSMT-28477/2014 within the HaBiT Project 7F14047and by the Conicyt-Fondecyt project "Detección automática del significado de los verbos del castellano por medio de patrones sintáctico-semánticos extraídos con estadística de corpus" (nr. 11140704, lead researcher: Irene Renau), which is funded by the Chilean Government.