Linking Verb Pattern Dictionaries of English and Spanish

Vít Baisa[1], **Sara Može**[2], **Irene Renau**[3]
[1]Masaryk University, Czech Republic, Brno
[2]University of Wolverhampton, United Kingdom
[3]Pontificia Universidad Católica de Valparaíso, Chile
xbaisa@fi.muni.cz,S.Moze@wlv.ac.uk,irene.renau@pucv.cl

Abstract

The paper presents the first step in the creation of a new multilingual and corpus-driven lexical resource by means of linking existing monolingual pattern dictionaries of English and Spanish verbs. The two dictionaries were compiled through Corpus Pattern Analysis (CPA) – an empirical procedure in corpus linguistics that associates word meaning with word use by means of analysis of phraseological patterns and collocations found in corpus data. This paper provides a first look into a number of practical issues arising from the task of linking corresponding patterns across languages via both manual and automatic procedures. In order to facilitate manual pattern linking, we implemented a heuristic-based algorithm to generate automatic suggestions for candidate verb pattern pairs, which obtained 80% precision. Our goal is to kick-start the development of a new resource for verbs that can be used by language learners, translators, editors and the research community alike.

Keywords: Corpus Pattern Analysis, Linked Data, Lexicography, Lexical Semantics, Bilingual resources

1. Introduction

This paper presents the results of a preliminary study in cross-linguistic pattern linking based on existing monolingual verb pattern dictionaries for English and Spanish, which are the outcomes of two separate research projects aiming to create freely available monolingual resources. The *Pattern Dictionary of English Verbs* (PDEV)[1] currently covers over 1,700 English verbs, whilst the *Pattern Dictionary of Spanish Verbs* (PDSV)[2] contains around 300 verbs (100 of which are currently available online). Both dictionaries were conceived as inventories of semantically motivated syntagmatic patterns, i.e. sentence structures and the semantic categorisation of the verb's arguments. Consider the example below:

1. [[Human | Institution]] avoids [[Eventuality]]
   Example: *The Government must avoid war.*

A common use of the verb 'avoid' has to do with a [[Human]] or an [[Institution]] trying to prevent an [[Eventuality]] from occurring. The capitalised words displayed between double square brackets are not lexical items, but 'semantic types', i.e. mnemonic labels that best describe the semantic features shared by the nouns that typically occur in a given argument slot. Syntactically, the verb occurs in a monotransitive construction. The observed sense of the verb 'avoid', i.e. *to prevent from occurring*, can only be activated by this specific combination of obligatory syntactic arguments (subject, direct object) and their corresponding semantic types. As a result, patterns allow us to unambiguously map word meanings onto their syntagmatic context, offering rich syntactic and semantic information about the verb's behaviour whilst providing exhaustive evidence from the corpus.

The present study represents our first attempt at linking equivalent verb patterns found in two or more languages. For instance, the pattern of the English verb *avoid* shown in 1 is equivalent to the following pattern exhibited by the Spanish verb *evitar*:

2. [[Human | Institution]] evitar [[Eventuality]]
   Example: *El Gobierno debe evitar la guerra.*

The two patterns are identical in that they are both transitive and use the same semantic categories ('semantic types') to describe their arguments. The meaning of both patterns is also the same. In this paper, we propose to match patterns English and Spanish verb patterns automatically by applying a heuristic-based algorithm that calculates the similarity between patterns. If successfully implemented, the algorithm will allow us to start building a bilingual lexical resource efficiently using PDEV and PDSV.

In recent years, multilingual lexical resources have been mushrooming all over the globe. Despite their coverage and suitability for different tasks and purposes, there resources have yet to successfully tackle the complexities of verb behaviour. A multilingual resource such as the one we propose here will have a number of potential applications in Natural Language Processing and language learning, and will provide empirically sound lexical data that can be used in theoretical and applied cross-linguistic studies.

The paper is structured as follows: Section 2. provides information on the theoretical and methodological background underpinning the proposed research and describes the two pattern dictionaries in more detail; section 2.2. features a short overview of related work in the field, and the following two sections focus the on the manual (Section 3.) and automatic (Section 4.) linking methods developed in this study. Finally, our plans for future are discussed in the Conclusion.

2. Background

2.1. Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) (Hanks, 2004a) is a corpus-driven technique that aims at mapping word meaning onto specific syntagmatic patterns exhibited by the target word

[1] www.pdev.org.uk
[2] www.verbario.com

in any type of text. Based on Theory of Norms and Exploitations (TNE) (Hanks, 2004b; Hanks, 2013), CPA aims at identifying patterns of normal usage (norms) and investigating the way the very same patterns are exploited creatively (exploitations) by means of in-depth, labour-intensive lexical analysis of corpus data. By doing so, it provides a window into the normal, every-day phraseology, which makes it particularly well-suited for both lexicographic and NLP tasks.

TNE and CPA are influenced by a large amount of cognitive, pragmatic and corpus linguistics studies interested in investigating how words interact in creating meaning and how this connection can be demonstrated using empirical data (see (Hanks, 2013) for a theoretical overview). CPA has been developed especially with lexicographical resources in mind, providing a solid alternative to 'classical', introspection-based analyses of meaning, which focus on words in isolation rather the way they behave in specific contexts. In CPA, meaning is pattern-based, not word-based. For instance, consider example 1 again:[3]

3. 1 [[Human | Institution]] avoids [[Eventuality]]
   2 [[Human | Animal]] avoids [[Physical Object]]

There are no syntactic differences between the two patterns - both are transitive, but the semantic types assigned to the subject and direct object do not match, hence the difference in meaning. More specifically, the first pattern refers to an action, process or state a human being or an institution tries to keep from occurring so that it does not affect them, whereas the second pattern refers to the reaction of a human or an animal trying not to physically interact with an object.

2.2. Multilingual Lexical Resources

The compilation of large, freely available multilingual lexical resources by means of linking pre-existing data has been gaining considerable traction in recent years, and justifiably so - once a monolingual resource has been created, it makes perfect sense to reuse and transform the data for different purposes. Bringing together compatible resources for different languages is particularly popular, as demonstrated by the existence of two major international projects in lexical analysis: WordNet, which allows researchers to connect and share their work through the Global Wordnet Association, see (Vossen, 2002),[4] and FrameNet (Fillmore and Baker, 2010),[5] whose infrastructure and data are being used by hundreds of researchers from all across the globe. PDEV and PDSV differ from the lexical resources developed in these two projects in that they do not share the same object of study: WordNet studies concepts linked to groups of verbs named *synsets*, FrameNet is centred around semantic frames, and CPA is corpus-driven and pattern-based. As a result, they can only be considered as complementary resources. As already pointed out, an important advantage of CPA is that it is particularly well-suited for verbs, as it allows researchers to perform fine-grained syntactic and semantic analysis of any verb's argument structure.

BabelNet[6], Omega Wiki[7], and Wiktionary[8] are other multilingual projects to be mentioned, which 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. Nevertheless, they lack an empirical basis, that is, they are not linked to corpus evidence.

Finally, in Language Learning, new tools are being created and offered online. A good example is Linguee[9], a tool that combines pairs of bilingual dictionaries in many languages with a parallel corpus showing the use of the target word in context. Another example is the Interactive Language Toolbox (Buyse and Verlinde, 2013), which was developed for second language learners. These are only two examples of how a bilingual or a multilingual dictionary can adapt to new technologies and users' needs and combine with non-lexicographical resources to provide an enhance user experience. Our proposal can be considered as a step in the same direction.

2.3. CPA Projects

The **Pattern Dictionary of English Verbs** (PDEV) is a publicly available resource developed in the DVC (Disambiguating Verbs by Collocation) project by Patrick Hanks' team at the University of Wolverhampton. The dictionary provides information on all the typical patterns associated with a verb, their definitions, and the corresponding corpus examples. For each verb, a corpus sample of 250 concordance lines is extracted from the British National Corpus (Leech, 1992), and tagged with pattern numbers using Sketch Engine (Kilgarriff et al., 2014). Depending on the semantic and syntactic complexity of the verb, the sample can be incrementally augmented to 500 or 1,000 concordance lines. Patterns are identified mainly through lexical analysis of corpus lines, complemented by the information found in the automatic collocations profile, *word sketches*[10], a feature available in the Sketch Engine, and are described using the CPA Editor (Baisa et al., 2015) and CPA's shallow ontology of semantic types[11]. Implicatures (pattern definitions) are written; register, domain, and idiom/phrasal verb labels are added, and links to FrameNet (Ruppenhofer et al., 2006) are created, 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 (Figure 1 shows PDEV entry for *harvest*). PDEV-lemon, a linked data implementation of PDEV is available (Maarouf et al., 2014).

The **Pattern Dictionary of Spanish Verbs** (PDSV) is currently being developed within the *Verbario* project at the Pontifical Catholic University of Valparaíso, Chile. The goal of the project is two-fold: 1) to perform manual analysis of the most frequent Spanish verbs using CPA as a working methodology, and 2) to develop and implement procedures aimed at automatizing the creation of new patterns

[3] For the full list of patterns, see: `http://pdev.org.uk/#browse?q=avoid;f=A;v=avoid`

[4] `globalwordnet.org`

[5] `framenet.icsi.berkeley.edu`

[6] `babelnet.org`

[7] `omegawiki.org`

[8] `wiktionary.org`

[9] `linguee.com`

[10] `en.wikipedia.org/wiki/Word_sketch`

[11] `pdev.org.uk/#onto`

(Nazar and Renau, in press). The project uses the same version of the CPA shallow ontology as the DVC project, as it proved to be equally valid for Spanish. In addition, the CPA ontology also serves as a top ontology in the creation of a new automatic taxonomy of Spanish nouns, which is being applied to the task of labelling verb arguments with semantic types (Nazar and Renau, 2016). PDSV is being built following the same guidelines as PDEV and with the continued support from the English team, which ensures compatibility between the two projects and ensuing lexical resources. The Spanish team uses the same database structure and corpus interface as the English team (i.e. the Sketch Engine), but they focus on high-frequency verbs (as opposed to the predominantly medium-frequency verbs currently contained in PDEV), and typically annotate slightly larger corpus samples (i.e. between 250 and 1,500, depending on the verb).
Finally, a considerable amount of work has been conducted in the application of CPA to Italian (Ježek et al., 2014), which resulted in the creation of a parallel Pattern Dictionary of Italian Verbs (PDIV).

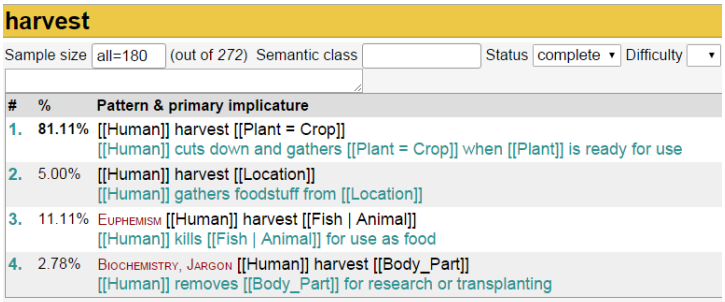

Figure 1: The dictionary entry for *harvest* in PDEV, as shown in the CPA Editor.

PDEV and PDSV are highly compatible in that they are being compiled using the same tools and methodology, making them perfect candidates for cross-linguistic pattern linking. In addition, CPA-based monolingual pattern dictionaries are developed independently of each other by different teams of lexicographers, which prevents dictionary data from being skewed due to possible interferences between languages. Corresponding pattern pairs in two or more languages can simply be linked to create a multilingual lexical resource based on their shared syntactic and semantic features. If successful, the proposed linking technique could make a significant contribution to the development of a new generation of multilingual lexical resources that focus explain meaning through patterns of real language use rather than abstract lists of word senses.

3. Manual Pattern Linking

In an effort to identify potential issues in the future, we decided to link patterns of a small subset of English and Spanish verbs. We selected 87 Spanish verbs with one or more English equivalents (126 in total), focusing on verb pairs such as *acusar* (accuse) and semantically equivalent groups of near synonyms such as *enfadar* (annoy/anger/infuriate/enrage). Pattern pairs identified through the manual linking procedure were later used as a gold standard in evaluation of the automatic linking task (Section 4.). Only verbs exhibiting up to 15 patterns were included in this pilot study, because highly polysemous verbs require specific strategies due to their grammatical complexity.
The study allowed us to identify the following methodological and practical issues that prevented us from finding full matches for all the patterns studied:

1. Both dictionaries differ significantly in terms of coverage: PDEV covers mainly low-to-middle frequent verbs, whereas PDSV contains middle-to-high frequent verbs. This reduces the number of potential matches; for instance, *golpear* (to hit) and to stab are often listed as translation equivalents in bilingual dictionaries despite the fact that their semantic overlap is very low.

2. The lack of full equivalence between languages, also known as anisomorphism (Yong and Peng, 2007). The following types of semantic anisomorphism were identified:

   (a) Lack of 1:1 correspondence: highly polysemous verbs typically exhibit a range of meanings and syntactic structures that differ significantly from their closest translation equivalents; in some cases, a pattern in a language might correspond to multiple patterns in the other language; e.g., for the previous example of *golpear*, a pattern such as '[[Human]] stabs (Physical Object 1) (at Physical Object 2)' could be considered equivalent to '[[Human]] golpear [[Physical Object]]', but the last one is too general to be matched to the English pattern.

   (b) Zero equivalence: some patterns simply do not have a corresponding pattern in the target language due to cultural, social, cognitive or pragmatic reasons. Idioms and other phraseological units are particularly problematic in that respect; e.g. the Spanish expression *sin comerlo ni beberlo* ('without being responsible for the damage caused to somebody'), which is listed as a pattern under the entry for *beber* (to drink), cannot be linked to any pattern for the verb *to drink*.

   (c) Syntactic differences: semantically equivalent pattern pairs often differ significantly in terms of their syntactic structure. A good example is the causative-inchoative alternation—a considerable portion of the verb pairs we studied showed that corresponding verbs often differ in the syntactic alternations they exhibit. For instance, the Spanish verb *agravar* exhibits both alternations, whereas its closest equivalent in English, *to aggravate*, can only be used in a causative construction.

4. Automatic Pattern Linking

To speed-up the labour-intensive procedure of manual linking, we decided to implement a heuristic-based algorithm

for automatic linking of pair candidates. Since the number of manually linked pattern pairs was very limited, it was not possible to train a machine learning system for the task. The small set of annotated manually pairs was used as a gold standard for evaluation of the method. Manual links are considered to be correct and the output of the automatic method will have to be constantly revised by lexicographers.

4.1. Algorithm

For each of the 490 Spanish patterns, we computed a similarity score for all its possible translations into English (i.e. verbs and their patterns, which resulted in a total of 5,067 Spanish-English pattern pairs). Candidate English patterns were then sorted by the score and the top pair was put forward as the best candidate for pattern linking.

The **similarity score** was computed by comparing pattern structures. Since this is a preliminary work, our analysis focused only on the three main syntactic arguments: subject, direct object and indirect object. An argument can have more than one semantic type associated with it, e.g. [[Human]] and [[Institution]] often occur together, as shown in Example 1. Whenever there was a non-empty intersection of semantic types in a given argument, 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]]). If, for instance, [[Event]] appears as the direct object in the Spanish pattern and [[Eventuality]] in its English counterpart, we can use the CPA ontology to get a partial match). Each hyponym or hypernym got score points based on the distance in the CPA ontology tree (the further apart they are located, the fewer score points they gain, measured in powers of 0.5). Scores for the three slots (subject, direct indirect object) were summed and the final score was assigned to the given pattern pair (cf. Table 1). All candidate pairs were sorted by the score and the top ranking pattern was returned.

Spanish	English	Scr	Comment
Entity|Eventuality	Human	1/8	Human < Animate < Physical_Object < Entity, distance = 3
Human	Human	1/2	Human is almost in all patterns so the score was only 0.5
Artifact	Eventuality	0	No relation in ontology

Table 1: Examples of ontology matches and the resulting scores (Scr) for pattern arguments.

The first column contains Spanish semantic types in a pattern argument. Since both PDEV and PDSV contain verbs with patterns containing semantic type Human in subject argument, the algorithm considers it as a weaker sign of equivalence. When two different semantic types S and E are in the same argument in a Spanish and an English pattern, CPA ontology (which is shared between PDEV and PDSV) is queried. If S is hypernym/hyponym of E or vice versa, the score is computed as 0.5^N where N is the distance in the ontology hierarchy (a tree in the case of CPA ontology).

Not all possible pattern pairs were considered, only patterns of equivalent English and Spanish verbs were taken into account. We have used a statistical English-Spanish dictionary derived from a parallel English-Spanish corpus. It is important to note that even if a verb in one language has more than one translation equivalent in the other language, the comparison of pattern structures should narrow the number of all possible pattern pairs—a pattern express one of possible meanings of a verb and it is reasonable to expect equivalent patterns to have the same or similar structure.

To evaluate the method, we created a random sample of 50 Spanish-English verb pairs. We excluded all cases in which a Spanish pattern cannot 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: 40 of the 50 pairs were correctly suggested as candidates and the rest was incorrect.

5. Conclusion

The paper presented the results of a pilot study on linking verb patterns across languages. Despite the fact that our work is currently at an early preliminary stage, the study clearly demonstrated the advantages of linking methodologically compatible, monolingual pattern dictionaries through a combination of both manual and automatic procedures. The algorithm developed for the task performed remarkably well considering the size of our gold standard dataset. There is plenty room for improvement—the manual task will have to be further refined, and the algorithm's performance improved by augmenting the size of the training data. Nonetheless, the work presented here will serve as a solid basis for the future development of the proposed methodology and ensuing lexical resource. Our immediate plans for the future include the creation of larger gold standard datasets of manually linked pattern pairs, as well as the adaptation of the software in a way that will allow lexicographers from different teams to manually specify links between two or more patterns contained in CPA-based pattern dictionaries. Our ultimate goal is to create a valuable, multilingual, corpus-driven lexical resource for verbs that reflects real language use and can therefore be used by language learners, language professionals (e.g. translators, editors) and the research community alike.

6. 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, 47. AH/J005940/1, 2012-2015] and 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.

7. References

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

Buyse, K. and Verlinde, S. (2013). Possible effects of free on line data driven lexicographic instruments on foreign language learning: The case of Linguee and the interactive language toolbox. In *Procedia: Social and Behavioral Sciences*, volume 95, pages 507–512. Elsevier BV.

Fillmore, C. J. and Baker, C. (2010). A frames approach to semantic analysis. *The Oxford Handbook of Linguistic Analysis*, pages 313–339.

Hanks, P. (2004a). Corpus Pattern Analysis. In *Euralex Proceedings*, volume 1, pages 87–98.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274.

Hanks, P. (2013). *Lexical Analysis: Norms and exploitations*. Mit Press.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., and Popescu, O. (2014). T-pas: A resource of corpus-derived types predicateargument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 26–31.

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

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

Maarouf, I. E., Bradbury, J., and Hanks, P. (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Nazar, R. and Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia.

Nazar, R. and Renau, I. (in press). A quantitative analysis of the semantics of verb-argument structures. In *Collocations and Other Lexical Combinations in Spanish. Theoretical, Lexicographical and Applied Perspectives.*, pages 92–108. Ohio University Press.

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

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

Yong, H. and Peng, J. (2007). *Bilingual Lexicography from a Communicative perspective*, volume 9. John Benjamins Publishing.