# Automatic Extraction of Lexical Patterns from Corpora

**Irene Renau, Rogelio Nazar**
Pontificia Universidad Católica de Valparaíso (Chile)
e-mail: irene.renau@pucv.cl, rogelio.nazar@pucv.cl

## Abstract

We present our first attempt to extract lexical patterns using corpus statistics. A pattern is a structure that combines syntactic and semantic features and is linked to a conventional meaning of a word. This means, for example, that the verb *to die* does not have intrinsic meanings, but potential meanings which are activated by the context: in 'His mother died when he was five', the meaning of the verb differs from 'His mother is dying to meet you', due to collocational restrictions and syntactic differences. With the automatic analysis of thousands of concordances per verb, we can make a first approach to the problem of detecting these structures in corpora, a very time-consuming task for lexicographers. The average precision is around 50%. The next step to increase precision is adding a dependency parser to the system and make adjustments to the automatic taxonomy we have created for semantic labeling.
**Keywords:** computational lexicography; lexical patterns; Spanish verbs; taxonomy.

## 1 Introduction

This paper presents the results of our first attempt to automatically extract patterns of the usage of words directly from corpus. Following Hank's (2004, 2013) approach, a pattern is a 'semantically motivated and recurrent piece of phraseology' (Ježek and Hanks 2010: 8). This notion is based on the idea that the meaning of a word is subordinated to the context in which the word is used. Depending on the context, the same word activates or inhibits one of its potential meanings (Hanks 2013: 73-75). This approach involves pragmatic, cognitive, textual and also linguistic considerations. Corpus linguistics has payed special attention to syntagmatic context of occurrence of the word (Firth 1957; Sinclair 2004, Hoey 2005, Hanks 2013). For instance, in the case of the verb *to follow*, we can observe that two of its patterns, even when they have the same syntactic structure, are nevertheless associated with different meanings (we take the examples from the *Pattern Dictionary of English Verbs,* PDEV, Hanks, in progress):

***Pattern 1:*** **Human 1** *or* **Animal 1** *or* **Vehicle 1 follows Human 2** *or* **Animal 2** *or* **Vehicle 2**
*Implicature:* Human 1 *or* Animal 1 *or* Vehicle 1 moves in the same direction as that selected by Human 2 *or* Animal 2 or the driver of Vehicle 2
*Example:* *They cluster around him in a dense shoal and* ***follow*** *him as he moves about.*

***Pattern 5:*** **Human** *or* **Institution follows Command** *or* **Rule** *or* **Plan** *or* **Document**
*Implicature:* Human *or* Institution acts in accordance with Command *or* Rule *or* Plan (expressed in Document)
*Example:* *Burton had once again* ***followed*** *the directive of an older man.*

Pattern 1 and 5 of the verb *to follow,* as analysed in PDEV, represent two phraseological structures involving syntactic and semantic features. In both cases, the verb is transitive, but the semantic

categories of the arguments vary: in the first pattern, for example, the subject is [[Human | Animal | Vehicle]], and in the second case is [[Human | Institution]]. These semantic types (Hanks 2013: 12-15) usually alternate, as happens in both subjects.
In other many cases, pattern distinction operates in the syntactic level as well:

***Pattern 1:*** Human dies ((Time Point)(Location)(Causation) (at Number *or* at **the age of** *or* at **birth** *or* **early age**))
*Implicature:* Human ceases to live ((Time_Point) ([Adv[Location) ([Adv[Causation) (at Number = Age in Years | at the age of | at birth *or* early age))
*Example:* *His mother **died** when he was five.*

***Pattern 8:*** Human is dying to-infinitive
*Implicature:* Human is very keen to/inf [verb]
*Example:* *My hon. Friend is **dying** to intervene.*

In these two patterns of the verb *to die,* one can observe that the semantic difference lays on syntactic structure (pattern 1 requires adverbials and pattern 8 requires an infinitive clause as a complement).
In sum, these examples show that two kinds of information are relevant when analysing the meaning of a verb: the syntax of the sentence where the verb is working as the predicate (summarised in subject, objects and complements structure) and the semantic categories of the arguments of the verb. To allow meaning differentiation, semantic categorisation requires a more fine-grained description than the one we can obtain with traditional semantic roles (such as 'agent' or 'patient'). Hence, a sufficiently large group of semantic types is needed, and this group needs to be hierarchically organised in a taxonomy. Nevertheless, it does not seem necessary to establish very detailed taxonomic categories when conducting semantic analysis for identifying lexical patterns of normal usage. For example, the semantic type [[Vehicle]] seems necessary because it is connected to a large number of actions such as *to follow* (see example above), *to crash, to stop, to turn* and so on. The semantic type [[Vehicle]] is necessary because it would not be parsimonious to create different categories for [[Jeep]], [[Bus]], [[Canoe]], etc.
In the following pages, we will describe a preliminary approach to the automatic collection of verb patterns such as the ones previously shown. The proposal is fully corpus-driven and it has been applied to Spanish. The algorithms, however, are based on statistical techniques and they can be applied to other languages too. Our proposal is attractive for lexicography because it helps to improve the quality and speed of work in the context of corpus-based dictionaries. Subsequent editing and correction is needed, but the automatic patterning can be the first step to prepare the manual tasks.

## 2 Methodology for conducting a statistically-based taxonomy of nouns

In a different paper, we explained how a combination of different algorithms can be used to build a taxonomy from corpus (Nazar and Renau 2016). As pointed out in the previous section, it is necessary to organise semantic types hierarchically to provide coherence. We use the CPA Ontology (Hanks, in progress), a shallow ontology of around 250 categories, to populate it automatically with approx. 35,000 Spanish nouns, using corpus statistics techniques. This allows us to automatically detect semantic types in text with an acceptable level of precision (see below). The full procedure to develop this taxonomy and to identify and categorise named entities is a complex endeavour in itself.

In the present paper we only describe how we applied this taxonomy for the specific goal of pattern building.
To summarize the procedure, these algorithms are of two types:

a) Some of them group together nouns which can be defined with the same hypernyms. For instance, one group may contain words such as *manzana* ('apple'), *fresa* ('strawberry'), *plátano* ('banana'), etc.
b) Other algorithms assign a hypernym to the groups that were created, i.e., they add the label *fruta* ('fruit') to the nouns exemplified in a).

By this procedure, nouns are interconnected between them and with CPA Ontology labels. For example, *fruta* is connected with the semantic type [[Plant Part]], which is one of the nodes in CPA Ontology. Once the 'ontological level' is achieved, the rest of the hierarchical structure is provided by the ontology.
As part of an ongoing project, our taxonomy is constantly growing. Of course, it is not realistic to try to populate a taxonomy with proper nouns due to their extreme diversity. Instead of that, we developed a mechanism for the detection of proper nouns and their categorisation as human, organisation and location, inspired on the work done in Named Entity Recognition and Categorization (Nadeau & Sekine, 2007; Grishman, 2012).
Figure 1 shows the noun *fresa* as an example of the option to get the hypernymy chains. The system offers 4 hypernymy chains (that is, four meaning candidates): *arbusto* ('bush'), *fruto* ('fruit'), *herramienta* ('tool') and *color* ('color'), the first one of them being only partially correct. Although it is correct that *fresa* is the name of the plant, this plant is not a bush.

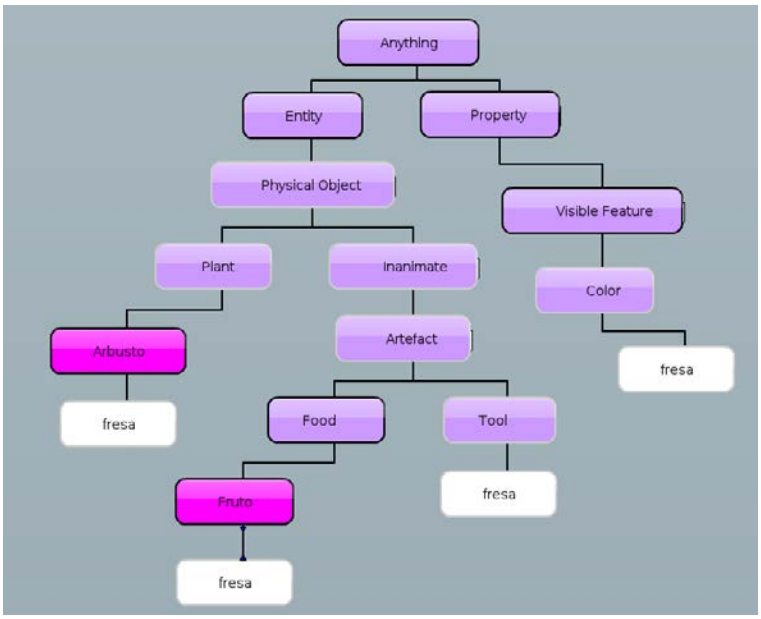


Figure 1: An example of a group of hypernymy chains for the Spanish word *fresa.*

Figure 2 shows an example of the hyponyms one can obtain when looking for the word *asiento* ('seat') in the taxonomy (it is only a fragment). The system offers 26 words which can be defined as types of 'seats', such as *diván, columpio, sillín,* etc. Some of them are incorrect, such as *asentamiento* ('the action of establishing in a place') or *respaldo* ('backrest'), a meronym.

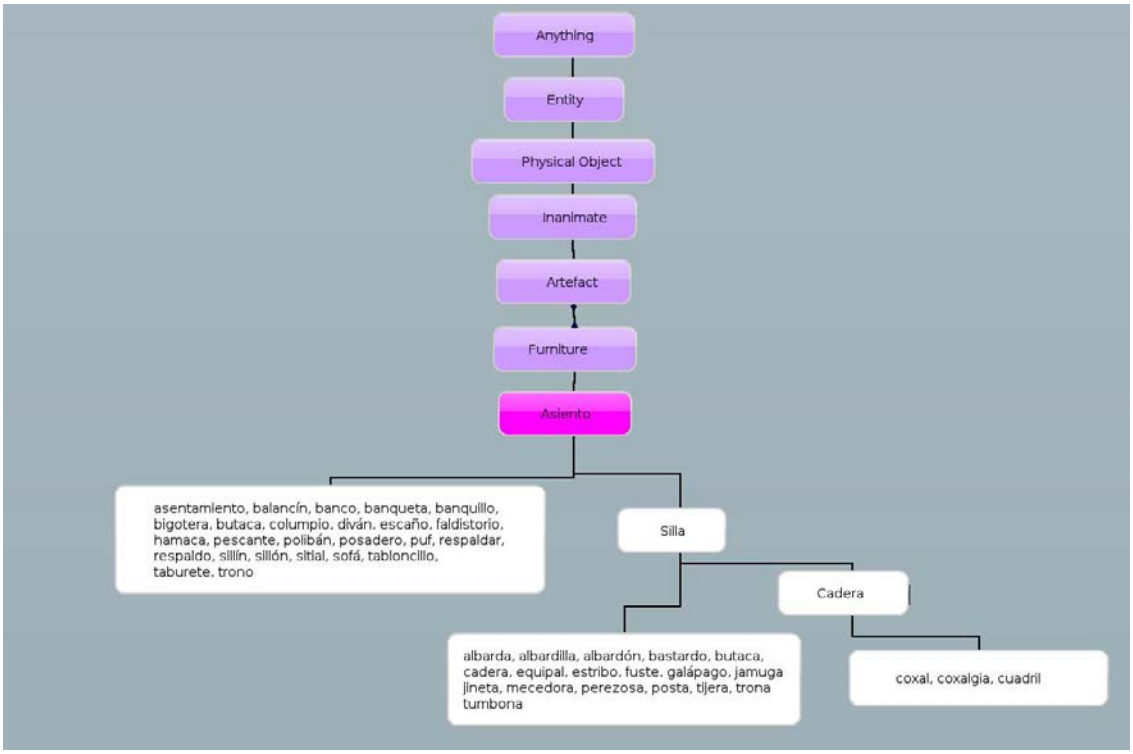


Figure 2: A fragment of the series of hyponyms of the Spanish noun *asiento* offered by the taxonomy.

A team of 6 advanced students in linguistics evaluated the methodology by taking 14 categories and observing if they were correctly matched with their hyponyms (in total, around 4,200) (Nazar and Renau, 2016). For example, we took a list of candidates to be *huesos* ('bones'), *mamíferos* ('mammals'), *colores* ('colours'), etc. and checked if the ISA relation was preserved (*coxis* 'coccyx' is a *bone, ballena* 'whale' is a *mammal,* etc.). We calculated overall precision as the ratio of correct chains over total chains and got around 77% precision (around 90% with high grade of certainty). Recall was calculated making an automatic comparison with Spanish WordNet 1.6 (Atserias, Villarejo and Rigau 2004). In our research we found out that only 25% of the nouns in WordNet were also included in our taxonomy, and only 32% of the nouns in our taxonomy were also included in WordNet. This comparison offers a glimpse of the size of Spanish vocabulary and also suggest the idea of merging of both resourses for future work.

## 3 Methodology for Pattern Extraction

Once with the taxonomy at hand, we used it to detect the structure of the lexical patterns, specifically the number and type of arguments. In this phase of the process, we wanted to evaluate precision without considering syntactic dependency parsing, that is, only taking into account co-occurrence statistics. We did this for simplicity, to reduce the need of external resources and to make the procedure as language-independent as possible.

The procedure for the extraction of the patterns was conducted in the following steps. For each verb *i* analysed (e.g. *acosar* 'to harass'), do:

1. Extract all sentences in the corpus where the analysed verb *i* occurs.

2. Replace in all possible cases the nouns that appear in the sentences by their corresponding semantic types using the taxonomy. E.g.: *problema* ('problem'), *dificultad* ('difficulty'), *peligro* ('danger') are replaced with the semantic type [[Eventuality]], and *hombre* ('man'), *mujer* ('woman'), *persona* ('person') with [[Human]].
3. Detect the prepositions used after *i*.
4. Build a data structure mapping each pattern with its example-sentences.
5. Detect patterns with pronominal uses of the verb *i* (in Spanish this is marked by the pronominal paradigm *me, te, se, nos, os, se* or by the enclitic).
6. Sort the patterns by decreasing order of frequency.

The procedure was applied to a sample of 100 Spanish verbs and can be consulted on the project's website (see section 2).

For the evaluation of our strategy, we took a group of verbs previously analysed by manual process as a gold standard (manual analysis can be also consulted online). These verbs were evaluated against the *Pattern Dictionary of Spanish Verbs* (PDSV), a lexical database containing phraseological patterns of the most frequent Spanish verbs. For the moment, a team of 6 lexicographers compiled around 300 verbs (only a sample is offered online), and we took a core of 30 randomly selected units for the evaluation. Criteria for the manual analysis and theoretical background have been described in Nazar and Renau (in press).

For the evaluation, we wanted to observe if combinations such as *subject + verb* or *verb + complement* in the automatic analysis were equivalent to the manually created patterns.

Among the examples of the automatic patterns, we find a sentence such as *Jones asustaba a los niños…* ('Jones scared the children…') correctly mapped to the pattern '[[Human]] *asustar a* [[Human]]'. That is, the two arguments of the verb are identified as human entities, and the preposition *a* is also registered. The pattern itself is the result of having found many similar sentences in the corpus.

As the maximum number of manual patterns in the sample taken for gold standard was 8, we arbitrarily took as a threshold the 10 most frequent automatic patterns for the evaluation. We only measured how many of the automatically produced patterns were correct, according to the previous manual work.

| verb | nr. of automatic patterns | nr. of correct patterns | precision (%) |
|---|---|---|---|
| abrir | 10 | 2 | 20.00 |
| afianzar | 10 | 1 | 10.00 |
| agrupar | 10 | 1 | 10.00 |
| alegrar | 10 | 6 | 60.00 |
| aproximar | 10 | 4 | 40.00 |
| avergonzar | 9 | 3 | 33.33 |
| cansar | 10 | 6 | 60.00 |
| colmar | 10 | 5 | 50.00 |
| conmover | 10 | 4 | 40.00 |
| cortar | 10 | 7 | 70.00 |
| cubrir | 10 | 8 | 80.00 |
| degradar | 10 | 5 | 50.00 |
| depurar | 10 | 4 | 40.00 |
| deslizar | 10 | 6 | 60.00 |
| disgustar | 10 | 1 | 10.00 |

| | | | |
|---|---|---|---|
| engrandecer | 10 | 2 | 20.00 |
| esperanzar | 10 | 9 | 90.00 |
| estremecer | 10 | 9 | 90.00 |
| exaltar | 10 | 4 | 40.00 |
| exasperar | 3 | 2 | 66.67 |
| generar | 10 | 4 | 40.00 |
| iluminar | 10 | 7 | 70.00 |
| instalar | 10 | 5 | 50.00 |
| llenar | 10 | 2 | 20.00 |
| motivar | 10 | 6 | 60.00 |
| Oxidar | 8 | 4 | 50.00 |
| precipitar | 10 | 4 | 40.00 |
| preocupar | 10 | 2 | 20.00 |
| reproducir | 10 | 1 | 10.00 |

Table 1: Summary of the results with a sample of verbs.

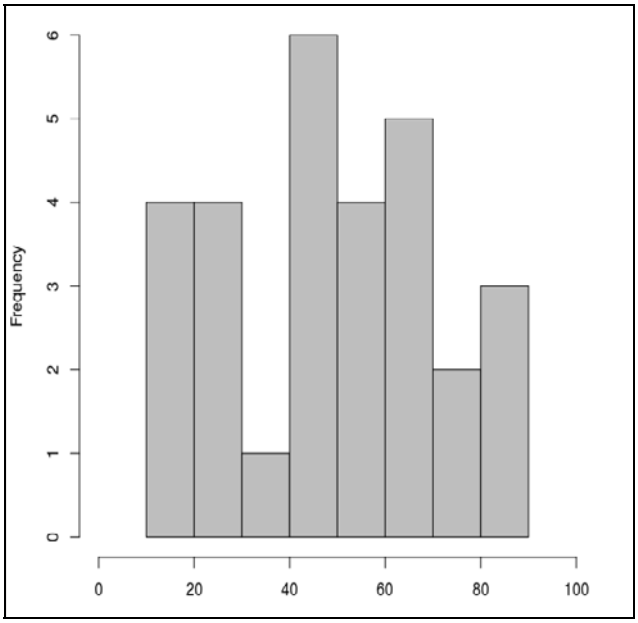


Figure 1: Histogram of precision in the results of the sample of verbs. The horizontal axis show percentage of precision and the vertical axis the number of times such figures appeared as a result of the automatic analysis of a verb.

Table 1 shows a summary of the results of the verbs of the sample. Figure 2 shows a histogram with the main tendencies. Precision seems to have a rather uniform distribution, but there is a mode between 40% and 50%. The main problems lay in the lack of specificity of syntactic structures –i.e., missing subjects, complements, prepositions, pronouns... We will address this problem by adding a dependency parser. With basic statistical analysis, the program succeeds in detecting the most frequent patterns of a verb (such as the one of *asustar* exemplified above), but it is not capable of offering correct proposals for less frequent patterns. It is also very common that results show the same pattern as if there were two or more different patterns –i.e. showing '[[Human]] *asustar a* [[Human]]' and '[[Human]] *asustar* [[Human]]' as separate patterns, when the fact is that they are the same. There are also errors related to polysemy, which leads to the incorrect semantic label of a noun

in a specific sentence. As it will be explained in the next section, all of these problems will be addressed in future work.

Sometimes, using the manual database as a gold standard played against recall because there were correct automatic patterns that were not found in the corpus sample that was analysed manually. This is in fact positive, because it demonstrates that the statistic procedure can improve manual results. We offer the following cases of the verb *degradar* ('to degrade, to deteriorate', etc.):

a) [[Human]] *degradar* [[Activity]]
*Example: ...pero no debe degradar su trabajo.* ('...but it should not degrade his/her work')

b) [[Human]] *degradar* [[Property]]
*Example: Debe ser difícil sin degradar la calidad del objeto digital.* ('it must be difficult without degrading the quality of digital object').

Both patterns are perfectly normal in Spanish and they were not found by lexicographers, probably due to the sample size (which varies between 200 and 1,500, depending on the grade of polysemy).

## 4 Conclusions and Future Work

This paper presented a brief summary of an ongoing project, including an evaluation of the preliminary results of the automatic pattern generation procedure. The results presented must be taken with caution, but we consider they are nevertheless promising and let us think that we could implement a useful tool for lexical analysis with lexicographic purposes.

In the next years we will develop three main lines of future work:

a) We will continue developing our work on the taxonomy induction system and we plan to include more algorithms to improve results.
b) We want to explore the inclusion of dependency parsing to determine if it helps to increase precision in the detection of lexical patterns.
c) We want to apply the same methodology for taxonomy induction and pattern extraction to languages different from Spanish. We are currently beginning to test the system with French and English.

As already mentioned, the final goal will be to develop a tool that will help lexicographers do their job faster and with less effort, which will naturally lead to operate with greater data samples and therefore with a magnified sense of objectivity for their analysis.

## 5 References

Atserias, J., Villarejo, L., Rigau, G. (2004). Spanish WordNet 1.6: Porting the Spanish WordNet across Princeton versions. In N. Calzolari, K. Choukri, T. Lino et al. (Eds.), *Proceedings of the Fourth International Conference on Language and Resources Evaluation (LREC),* pp. 161-164.

Bullinaria, J.A. (2008). Semantic categorization using simple word co-occurrence statistics. ESSLLI Workshop on Distributional Lexical Semantics.

Firth, J.R. (1957). Papers in Linguistics. London: Oxford University Press.

Grishman, R. (2012). Information Extraction: Capabilities and Challenges. Notes prepared for the 2012 International Winter School in Language and Speech Technologies. Rovira i Virgili University. Tarragona, Spain .

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

Hanks, P. (2004). Corpus pattern analysis. *Proceedings of EURALEX*, pp. 87-97, Lorient, France.

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

Hanks, P. (Ed.) (in progress). *Pattern Dictionary of English Verbs.* URL: http://pdev.org.uk [29/4/2016].

Hanks, P. (in progress). CPA Ontology. URL: http://pdev.org.uk/#onto [29/4/2016].

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146-162.

Hoey, M. (2005). *Lexical priming: a new theory of words and language,* London: Routledge.

Nadeau, D., Sekine, S. (2007). A survey of named entity recognition and classification. Journal of Linguisticae Investigationes 30:1.

Nazar, R.; Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. In N. Calzolari et al (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association (ELRA), May 2016.

Nazar, R., Renau, I. (in press). A Quantitative Analysis of the Semantics Of Verb-Argument Structures. In S. Torner and E. Bernal (Eds.), *Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches.* New York: Routledge.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse.* New York: Routledge.

## Acknowledgements

We want to thank Patrick Hanks for his continuous support to our work.

This work is being carried out with support of the following Conicyt-Fondecyt projects funded by the Chilean Government: 'Detección automática del significado de los verbos del castellano por medio de patrones sintáctico-semánticos extraídos con estadística de corpus', nr. 11140704 (lead researcher: I. Renau) and 'Inducción automática de taxonomías de sustantivos generales y especializados a partir de corpus textuales desde el enfoque de la lingüística cuantitativa', nr. 11140686 (lead researcher: R. Nazar).