Experiments in taxonomy induction in Spanish and French

Irene Renau
Pontificia Universidad
Católica de Valparaíso
irene.renau@pucv.cl

Rogelio Nazar
Pontificia Universidad
Católica de Valparaíso
rogelio.nazar@pucv.cl

Rafael Marín
Centre National de
la Recherche Scientifique
(Lille)
rafael.marin@univ-lille3.fr

Abstract

We present an ongoing project on taxonomy induction of nouns in Spanish and French. Experiments were first run in Spanish and, in this paper, we replicate the same method for French. Lexical taxonomies connect nouns following the IS-A structure: *árbol* ('tree') is a *planta* ('planta') is a *ser vivo* ('living being') is a *objeto físico* ('physical object'). In our proposal, we use a handmade shallow ontology of around 250 nodes and link every noun to one of these nodes. We use a set of algorithms based on corpus statistics techniques to build the hypernym-hyponym relations. As a result, any noun of Spanish or French can be linked to the taxonomy. Evaluation shows 60-90% precision, taking into account the best measures. At this stage of the process, our taxonomies can be already used for several NLP tasks such as semantic tagging of corpora, population of other taxonomies such as WordNet or applications in terminology. All the algorithms and a demo interface are available at <http://www.tecling.com/kind>.

1 Introduction

The present paper[1] describes a methodology for taxonomy induction in Spanish and French, using a combination of algorithms based on different quantitative approaches. At this stage of the project, we start with nouns because they are a central part-of-speech for conceptual categories. In our proposal, the major part of the algorithms receive raw corpus data as input, and as a result of all the process we obtain a taxonomic structure as output, linking each noun with its hypernym and building a hypernym chain. Previous results, as well as the algorithms used for the experiments and other material, are already published in http://www.tecling.com, a web page which is updated as we progress in the project.

From the lexical point of view, a taxonomy can be described as a structure of hypernymy relations, the so-called "IS A relations", e.g. *un martillo ES UNA herramienta ES UN artefacto ES UN objeto físico* ('a hammer IS A tool IS AN artifact IS a physical object'). Lexical taxonomies can contain other types of lexical relations such as synonymy or meronymy, as well as different parts-of-speech (verbs, nouns, adjectives, etc.). They are useful for a variety of tasks in natural language processing, as they organise raw linguistic data such as corpora. For example, they play an important role in corpus-based terminology and lexicography, as part of the process for automatising vocabulary extraction, creation of dictionaries, search for new terms, among other typical tasks in these areas.

Our approach in this project is mainly quantitative in order to facilitate the replication of the same experiments in different languages, as we do in the present paper for Spanish and French. Other languages will be included to the project as we progress.

[1] This paper received support of the Fondecyt Program (Conicyt, Chilean Government), Project nr. 11140704 and from Programa de Cooperación Científica Ecos-Conicyt, Project nr. C16H02.

We have been able to reduce the error rates of the procedure by using different algorithms combined, using a decision algorithm to decide via a voting system. Not all of the individual algorithms we use are new, but the novelty of the proposal lays on the way these algorithms are connected in a unified system.

In the following pages, we make a brief account of the state of the art in automatic taxonomy induction (section 2), we present our methodology (section 3), the results of the experiment conducted both for Spanish and French (section 4) and some conclusions and perspectives of future work (section 5).

2 Taxonomy induction: state of the art

There are countless ontologies or taxonomies used in a broad range of disciplines or professional areas, and the vast majority of these resources have been manually compiled by experts. For example, Cyc (Lenat, 1995) is an ontology for the general knowledge used for a variety of tasks in artificial intelligence; WordNet (Miller, 1995) and EuroWordNet (Vossen, 2004) are well-known taxonomies originally built by psychologists and linguists and widely used in natural language processing; and the CPA Ontology (Hanks, 2017a) is a shallow ontology used for semantic annotation of corpus data in a lexicographic project, the *Pattern Dictionary of English Verbs*, PDEV (Hanks, 2017b).

Manual resources have high precision, but they deal with different problems as well, the most important of them being how to update the resource without counting with a large team of trained experts working constantly on it. Initiatives such as the Observatory of Neology show that one can find new words and meanings almost in any copy of a newspaper, and that lexical and semantic change is the natural state of vocabulary. The same could be said about terminology, using scientific papers as source of information. For that reason, computational linguistics has been interested in the problem of taxonomy induction for decades.

First methods, conducted during the 70s and 80s, used computer-based dictionaries sources of taxonomic relations between the *definiendum* or hyponym and the *definiens* or hypernym. Hyernymy relations were extracted from dictionaries with rule-based methods (Calzolari, 1977; Amsler, 1981; Chodorow et al., 1985; Alshawi, 1989; Fox et al., 1988; Guthrie et al., 1990, among others). The advantage of these proposals was that they used reliable sources which can be considered already partially structured, as dictionaries work as "implicit taxonomies". However, these methods inherited the problems of lexicographic material, especially regarding the updating of the data but also in relation to the reliability of the data, because many dictionaries are not corpus-based even today.

Hearst (1992) proposed another strategy based on corpus linguistics, consisting of extracting definitional patterns from texts. For example, in a context such as "apples and other types of fruit", the pattern is "X and other types of Y", being X the hyponym and Y the hypernym. The strategy has been used in many studies (Rydin, 2002; Snow et al., 2006; Potrich and Pianta, 2008; Auger and Barrière, 2008; Aussenac-Gilles and Jacques, 2008, among others). This method is based on real data and facilitates the updating of information. However, it depends on a large amount of definitional rules, manually detected and compiled. Furthermore, these rules are language-dependent, which adds a difficulty to multilingual resources and in terms of replicability.

A third strategy consists of applying quantitative methods to taxonomy induction. Two main views can be outlined: on the one hand, many studies have shown interest in finding co-hyponym relations; that is, groups of words that are defined with the same hypernym, e.g. types of fruit, cheese, arms, emotions… (Grefenstette, 1994; Schütze and Pedersen, 1997; Lin, 1998; Alfonseca and Manandhar, 2002; Bullinaria, 2008). These words are said to be paradigmaticaly related, meaning that they tend to occur in similar syntagmatic contexts. Therefore, they are expected to share semantic features.

Another strategy consists of connecting hypernyms with their hyponyms through their asymmetric relationship when finding them in corpus: e.g. in a hypernym-hyponym pair such as *herramienta-martillo* ('tool-hammer'), it is more likely that *martillo* will appear in sentences with *herramienta* than vice versa, because *herramienta* can be used with other co-hyponyms of *martillo* such as *destornillador, llave, alicates* ('screwdriver, wrench, pliers'), etc. (Nazar et al., 2012). Also, as we do in this paper, Santus et al. (2014) also connect both tasks to create hypernymy chains using a combination of measures based on distributional semantics. Quantitative methods have the lack of precision as a potential problem, but the lack of certainty is compensated by the large amount of linguistic data. For that reason, this approach has become more popular and competitive since larger corpora have been available. Furthermore, being language-independent, they can be easily replicated and used to create multilingual resources.

3 Methodology

The methodology used for our experiments used the two quantitative approaches that were described in the previous section, combined. The general strategy consists of using an already created shallow ontology to build the top nodes of the taxonomy, which will be populated with the hypernymy chains, the latter step being the central part of the procedure. Spanish nouns are connected between them and also to the ontology nodes, building a hierarchical structure that includes the major part of the Spanish nouns, and any new noun can be processed and included in the taxonomy. The same procedure is applied to French. Both Spanish and French taxonomies are not connected at this stage of the project, but that is a task we are preparing for future work.

3.1 Materials

We used the CPA Ontology (Hanks, 2017a) to build the top nodes of the taxonomy. CPA Ontology is a shallow ontology of around 250 very general semantic types such as [[Process]], [[Action]], [[Physical Object]], etc. They do not include specialised information and many of them can be considered semantic primes (Wierzbicka, 1996), that is, concepts that cannot be defined with other concepts. For that reason, we consider the CPA Ontology as valid for any European language despite being originally created for English. Conversely, it would be not appropriate to use it when working with languages connected with very different cultures, such as the American indigenous languages or others.

Following the logic of using the CPA Ontology for the top nodes and leaving the automatic part for the most specific words, in our system the connection *roble* > *árbol* > *planta* ('oak > tree > plant') is automatic, but the connection *planta* > *objeto físico* > *entidad* > ('plant > physical object > entity') is part of the CPA Ontology. This way, most of the links have to be created automatically, but not in the case of the most general ones. Of course, the population of this shallow ontology (the process of connecting the nouns to the CPA's semantic types) is also automatic. This connection is triggered when a hypernym candidate is formally identical to some CPA semantic type.

The CPA is used only as a basic structure –it contains only 250 nodes which can be easily and even automatically translated to other languages. It has to be clarified as well that we can use any other ontology or taxonomy for the same purpose, and even the methodology can be applied to populate already existing resources such as WordNet. For example, we are starting to work with specialised vocabulary of Psychiatry, and for that purpose we are using a different ontology, also very general and with only 50 basic nodes.

Concerning the corpora, for algorithm 1 we used a lexicographic corpus which was necessary for one of the steps of the methodology. This corpus, consisting of noun definitions taken from online dictionaries, is a text file that has, in each row, nouns next to their definitions and, separated by a tab, different definitions for the same noun[2]. These definitions are used as plain text corpus, without metadata. For the algorithms 2 and 3, we used plain text extracted from Wikipedia, around 900 million words, without metadata or any kind of tagging. We used this corpus because it is big and open access, but the same method can be applied to any corpus with a similar or a larger size.

[2] We are preparing a different paper in which we explain our method to acquire, for any input noun, a set of definitions from the web.

3.2 Methods

3.2.1 Algorithm 1: analysis of *definiens-definiendum* co-occurrence

This algorithm analyses the lexicographic corpus to find hypernym-hyponym connections. Lexicographic entries are treated as plain text and all the text of the entries of all dictionaries sharing the same headword are grouped together in a sub-corpus, e.g. we group all the dictionary entries of *martillo* ('hammer'), obtaining a small set of raw text containing all the definitions of the different meanings of the word and even noisy information such as grammatical notes, etymology or abbreviations. The algorithm counts the number of times that a noun (the hypernym) co-occurs with nouns in the definitions (hypernym candidates). We assume that the noun which is more frequently used in the definitions of the different dictionaries in a specific entry is the hypernym, or hypernyms if the word is polysemous. For example, most of the dictionaries define *martillo* as *herramienta* ('tool'), which allows to create an IS A structure such as *martillo ES UNA herramienta* ('a hammer IS A tool').

The algorithm creates a list of candidates that correspond with the meaning(s) of the noun, eg. *herramienta, hueso, pieza, persona* ('tool, bone, piece, person'), etc. After the application of the rest of the algorithms, the results are confirmed or dismissed.

3.2.2 Algorithm 2: analysis of the asymmetric syntagmatic association

This algorithm uses the Wikipedia corpus to calculate the number of times that a target noun co-occurs with other nouns, then it calculates the number of times that one of these nouns co-occurs with the former noun. Based on the idea of asymmetric association between the hypernym and the hyponym, it is postulated that the hyponym tends to appear in the same sentences as its hypernym, but not the the other way around. We calculated these relations with directed charts that represent the co-occurrence relations of each word, in first and second degree. Figure 1 shows a graph representing these asymmetric relations found in corpus.

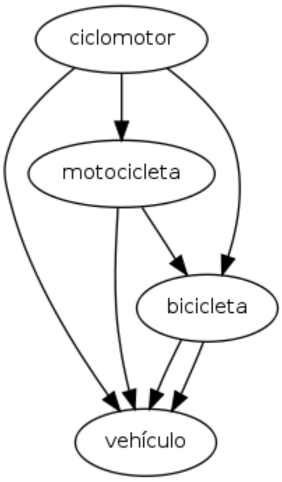

Figure 1: Example of a co-occurrence graph depicting the asymmetric relations between *ciclomotor* ('mopped') and its correct hypernym *vehículo* ('vehicle').

As observed in the graph, for the term *ciclomotor* ('moped'), the first-degree analysis points that it co-occurs with *motocicleta* ('motorbike') and *bicicleta* ('bicycle'). From this new analysis, we can observe that these two words appear in the same contexts that *vehículo* ('vehicle'), but this last term does not appear next to *ciclomotor, motocicleta* or *bicicleta*. These asymmetric relations are the ones considered hypernym clues. As a consequence, it can be concluded that "*ciclomotor* is a type of *vehículo*", simply because in this graph this is the node with the largest number of incoming arrows.

As in the case of the algorithm 1, here we also obtain hypernymy relations, in this case using a general corpus and with a different strategy. This pair of algorithms are necessary to build the taxonomical structure.

3.2.3 Algorithm 3: calculation of distributional similarity

As the algorithm 2, this algorithm also uses the Wikipedia corpus, but in this case to group different nouns sharing the same semantic type according to their distributional similarity. For example, the lexical items that refer to types of drinks, such as *café*, *vino*, *cerveza*, *té* ('coffee, wine, beer, tea', etc.) will show a tendency to appear in the same sentences with the same group of other units, such as *vaso*, *botella*, *beber*, (glass, bottle, drink, etc.). Therefore, for *café* there are bigrams such as *mucho café* ('a lot of coffee'), *buen café* ('good coffee'), *café ardiente* ('very hot coffee'), *café robusta* ('robusta coffee'), *tomar café* ('drink coffee'), etc. Each analysed word is associated with the lexical items co-occurring with it, and this association is represented as a word-vector, e.g. *café* = {*mucho, buen, ardiente, robusta, tomar...*}.

Once all analysed words are represented as vectors, the algorithm compares all of them against each other applying a similarity measure –the Jaccard coefficient– which calculates the degree of overlapping between vectors. As a result, we obtain groups of co-hyponyms, that is, words that can be defined with the same noun. This content is used to populate the labels that we previously obtain with algorithms 1 and 2. For example, if these algorithms established that *café* 'coffee' is a type of *bebida* 'drink', then every co-hyponym of *café* (such as *vino, cerveza, té...* in the previous example) will also be a type of 'drink'.

3.2.4 Algorithm 4: calculation of lexical and morphological similarity

This algorithm learns from the association between the lexical and formal features of the words with the conceptual categories they belong to. Unlike the previous algorithms, this particular one is not corpus-based. Instead, it only uses formal, non-linguistic information (such as components of the word defined as sequences of up to five letters at the beginning or end of each word). This way, if the system finds a lexical unit which cannot be found in corpus or if it is too infrequent to be analysed with the previous algorithms, then it will attempt to categorise whit unit using these formal features, in a process we term "analogical inference", because it learns from the categorisations conducted by the other algorithms.

In the lexical level, for example, it is possible to assume that if the previous algorithms have classified words such as *enfermedad celíaca* ('celiac disease') or *enfermedad pulmonar* ('lung disease') as hyponyms of *enfermedad* ('disease'), then via this analogical inference algorithm our system will classify a rarely used term such as *enfermedad de Knights* ('Knights' disease') as *enfermedad.* Also in the case of infrequent words such as *diverticulitis* ('diverticulitis'), the algorithm is able to infer that this word belong to the same group as other more frequent words, such as *apendicitis, laringitis* or *meningitis* ('appendicitis, laryngitis, meningitis'), because they share the same ending. This algorithm provides more flexibility and power of generalization to the system, since it implies a learning process that is conducted simultaneously to the analysis.

3.2.5 Algorithm 5: integration of methods

This final algorithm is in charge of the task of combining the information produced previous ones. Some of the previous algorithms collaborate and others reinforce the tasks already conducted. This integration is organised by a weighted voting procedure, considering the output generated by each of the algorithms presented above. It is weighted because algorithm 2 has twice the weight in this decision. In the event that a target word is found as a hyponym of both a direct parent and a grandparent, then the only criterion to decide between the two is the one that has been more frequently voted by the algorithms.

Furthermore, each decision will have attached a degree of certainty. For instance, if for an input noun there are more than two algorithms that coincide in placing such noun under a certain category, then the hypernymy link is presented with a high degree of certainty. If, instead, only two algorithms coincide in this, then such link only has a low degree of certainty. If only one algorithm is proposing this link, the proposal is ignored.

4 Results and evaluation

Results are shown as a list of candidates, each one taking the form of a hypernymy chain. The following is an example of such chains:

árbol > *planta* > *entidad* > *todo*

Here, the target word is the Spanish noun *árbol* ('tree'), which is automatically linked to its hypernym *planta* ('plant'). Then, the rest of the links (*planta* > *entidad* > *todo* 'plant > entity > everything') belong to the original structure of the CPA Ontology. Figures 2 and 3 show a graphic representation of the hypernymy chains for Spanish and French respectively.

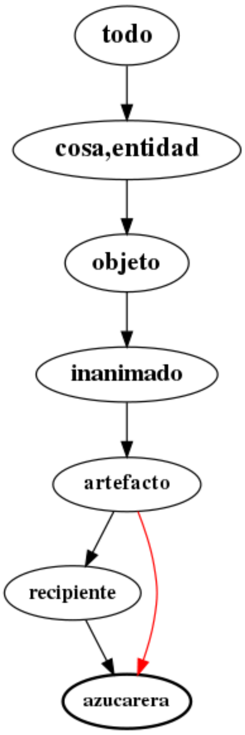

Figure 2: Result for Spanish noun *azucarera* ('sugar bowl')

In the example of figure 2, *azucarera* ('sugar bowl') is automatically linked to *recipiente* ('container') and *artefacto* ('artifact'), both semantic types of the CPA Ontology such as the rest of the nodes over them. Both links are correct, with different levels of semantic specification. In figure 3, the French word *bicyclette* ('bike') is also correctly linked to *véhicule roulant* ('vehicle with weels') and *véhicule* ('vehicle'), but the link to *roue* ('weel') is incorrect (it is actually a meronym). There are other correct and incorrect links in the structure shown in the figure, which is only a part of the whole net, e.g. the hyponyms linked to *bicyclette* are correct in the case of *ciclo-taxi* ('cycle taxi') but incorrect in the rest of the cases.

Regarding evaluation, we made a random sample of 100 nouns for each language and manually checked if the algorithm assigned hypernyms for each of them correctly. The sample is not stratified by frequency, which is detrimental for performance as most of the randomly selected words are infrequent. However, we leave for future work the development of an improved evaluation method.

Both for Spanish and French, criteria for precision consisted of considering as correct only those results with links that corresponded to a hypernym-hyponym relation, that is, when the target word could

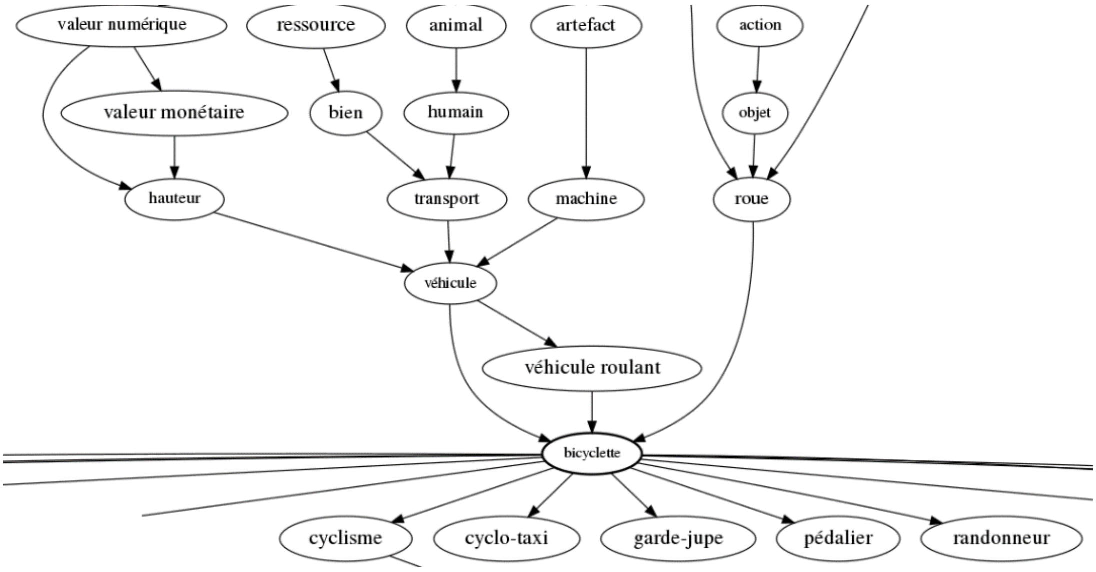

Figure 3: Result for French noun *bicyclette* (bike)

Table 1: Percentages of precision in the two languages by degree of certainty and rank of the candidate.

	French		**Spanish**	
Rank	**High certainty**	**All**	**High certainty**	**All**
1	60	51	54	46
2	76	65	74	65
3	83	70	78	68
4	90	74	78	71

be correctly linked to the upper node with the expression. In other words, we say there is a hypernym link between nouns X and Y if we can hold a statement such as "X is a type of Y", as in a "*bicyclette* is a type of *véhicule*". The rest of cases were considered incorrect. For instance, we marked as incorrect results such as "*termosifón* ('thermosyphon') is a type of *temperature* ('temperature')" for Spanish, or "*instructeur* ('instructor') is a type of *instruction* ('instruction')" for French. At this stage of our project, we did not calculate recall. Recall could in principle be defined as the number of senses detected per word over the total number of senses that actually exist for such word. We observed, however, that in the majority of the cases the system was only able to detect the most frequent meanings.

Precision was evaluated taking into account each position of the ranking and the degree of certainty of the algorithm. The rank of a candidate is given by the integration voting algorithm, thus the best candidate will be in the first position of the rank. We only considered the first 4 positions in ranking. Table 1 shows the intersection of results indicating high probability of success in each ranking position. If we only consider results ranked in the first position and with high degree of certainty, we obtain 60% precision in the French taxonomy and a 54% in the Spanish taxonomy. If we ignore the certainty level, results in first position drop to to 51% in French and a 46% in Spanish. Percentages of precision increase as we consider more positions in the ranking because then the system has more opportunities to find a correct hypernym.

The error analysis indicates that the major part of the errors are cases of semantic relations other than hypernymy. Typically, we found meronymy relations but also synonymy, co-hyponymy and even hyponymy. For example, in Spanish, the relation *aposento* > *edificio* ('chamber > building') corresponds to meronymy (*aposento* IS A PART OF *edificio*), and the same happens in French with *glacière* > *eau* ('glacier > water'). Also in the case of French, for the noun *produit* ('product'), one of the candidates for hypernym is actually a hyponym: *oeuvre d'art* ('piece of work'). Also incorrect is a result such as *copa* > *vaso* ('cup > glass') for Spanish, because the target word and the candidate are co-hyponyms.

Some of the errors are also due to interferences with the lexicographical marks coming from algorithm 1, such as in the case of the Spanish noun *pubis* ('pubis'), for which the candidate is *plural* ('plural'), due to the fact that many dictionaries indicate that the plural of this word is irregular. Problems regarding more general aspects of the methodology are related to the fact that the system does not distinguish between different candidates and different meanings at this stage of the project. Thus, for example, for a Spanish noun such as *taza* ('cup'), the system offers 4 candidates: *artefacto, vasija, recipiente* and *leche* ('artifact, vessel, vessel' and 'milk'). The first three candidates are correct, but they belong to the same meaning of the word, that is, all of them could be considered equivalent hypernyms. In the case of *artefacto*, the hypernym is the most general one, but it is correct because a cup is a type of physical object created by humans. The other two correct candidates are synonyms and, thus, equivalent and correct hypernyms, being *vasija* the old-fashion word and *recipiente* the modern one. Working on improving all these problems is part of our future work with the taxonomy project.

5 Conclusions and future work

In this paper, we have explained a methodology for creating a taxonomy based on a series of algorithms using different statistical approaches. Results shown in the previous section allow us to observe the advantages of the methodology, which connects a large number of vocabulary units via a corpus-driven analysis. The percentages of precision are still in need of improvement, but they are good enough to use the taxonomy for corpus semantic tagging and other NLP tasks. Renau and Nazar (2017), for instance, used these algorithm to tag arguments in order to study the semantics of verbs.

There are still a number of problems to be addressed in future work. For example, we are already testing the same method for specialised vocabulary, using a terminological ontology instead of the CPA Ontology, which was created for the analysis of general vocabulary. We are now working on different options to address the problems of polysemy, which are also an important source of problems in our taxonomy. In general, a more precise work is needed regarding evaluation and error analysis.

Another problem left for future work is to develop some strategy for the cases when a target word is found as a hyponym of both a direct parent and a grandparent. Now we only use the voting criterion, but a more sophisticated solution should be found, as some sort of reasoner which would be able to detect that both competing hypernyms are themselves a hyponym-hypernym pair. Similarly, the creation of a multilingual resource which could line up the taxonomies of Spanish, French and possibly other languages created independently is also left for future work. This alignment would be made by the extraction of bilingual vocabularies using parallel and comparable corpora.

References

Alfonseca, E. and S. Manandhar (2002). Extending a lexical ontology by a combination of distributional semantics signatures. In *International Conference on Knowledge Engineering and Knowledge Management*, pp. 1–7. Springer.

Alshawi, H. (1989). Computational lexicography for natural language processing. Chapter Analysing the Dictionary Definitions, pp. 153–69. White Plains, NY: Longman Publishing Group.

Amsler, R. A. (1981). A taxonomy for english nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pp. 133–38. Association for Computational Linguistics.

Auger, A. and C. Barrière (2008). Pattern-based approaches to semantic relation extraction special issue of terminology. *Terminology 14*(1), 1–19.

Aussenac-Gilles, N. and M.-P. Jacques (2008). Designing and evaluating patterns for relation acquisition from texts with caméléon. *Terminology 14*(1), 45–73.

Bullinaria, J. A. (2008). Semantic categorization using simple word co-occurrence statistics. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pp. 1–8.

Calzolari, N. (1977). An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie Paris 31*(2), 118–28.

Chodorow, M. S., R. J. Byrd, and G. E. Heidorn (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pp. 299–304. Association for Computational Linguistics.

Fox, E. A., J. T. Nutter, T. Ahlswede, M. Evens, and J. Markowitz (1988). Building a large thesaurus for information retrieval. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 101–8. Association for Computational Linguistics.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Guthrie, L., B. Slator, Y. Wilks, and R. Bruce (1990). Is there content in empty heads? In *Proc. of the 13th International Conference on Computational Linguistics, COLING'90 (Helsinki, Finland)*, pp. 138–143.

Hanks, P. (2017a). CPA ontology. http://www.pdev.org.uk/#onto. [last access: 31/8/2017].

Hanks, P. (2017b). Pattern dictionary of english verbs. http://www.pdev.org.uk/. [last access: 31/8/2017].

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pp. 539–45. Association for Computational Linguistics.

Lenat, D. (1995). Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM 38*(11), 33–38.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational linguistics-Volume 2*, pp. 768–74. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM 38*(11), 39–41.

Nazar, R., J. Vivaldi, and L. Wanner (2012). Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del Lenguaje Natural 49*, 67–74.

Potrich, A. and E. Pianta (2008, May). L-isa: Learning domain specific isa-relations from the web. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008).*

Renau, I. and R. Nazar (2017). Verbos in contexto: una propuesta para la detección automática de patrones léxicos en corpus. In I. Sariego López, J. G. Cuadrado, and C. G. Escribano (Eds.), *El diccionario en la encrucijada: de la sintaxis y la cultura al desafío digital*, pp. 879– 897. Santander: AELEX.

Rydin, S. (2002, July). Building a hyponymy lexicon with hierarchical structure. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, Philadelphia, Pennsylvania, USA, pp. 26–33. Association for Computational Linguistics.

Santus, E., A. Lenci, Q. Lu, and S. Shulte im Walde (2014). Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 38–42.

Schütze, H. and J. O. Pedersen (1997, May). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management 33*(3), 307–18.

Snow, R., D. Jurafsky, and A. Y. Ng (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 801–8. Association for Computational Linguistics.

Vossen, P. (2004). Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography 17*(2), 161–173.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.