Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 83-90 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

Hypernymy relations from definiens-definiendum co-occurrence in multiple dictionary definitions*

Relaciones de hiperonimia a partir de la coocurrencia definiens-definiendum en múltiples definiciones de diccionario

Irene Renau **Rogelio Nazar**
University Institute for Applied Linguistics
Universitat Pompeu Fabra
Roc Boronat 138
08018 Barcelona, Spain
{irene.renau;rogelio.nazar}@upf.edu

Resumen: Presentamos una metodología basada en estadísticas de coocurrencia entre *definiens* y *definiendum* con el fin de extraer relaciones hiperonímicas de un corpus lexicográfico, como parte de un proyecto más extenso dedicado a la creación de una ontología general de nombres aplicada al estudio de las relaciones predicado-argumento. La idea de la presente propuesta es hacer emerger las relaciones de hiperonimia mediante la combinación de distintas fuentes lexicográficas. Encontramos que los hiperónimos de una palabra son los que aparecen con más frecuencia en las definiciones de esa palabra en diccionarios y que, del mismo modo, sus hipónimos suelen ser los que contienen frecuentes menciones a esta palabra en sus definiciones. Esto crea una asociación estadística entre palabras y permite estructurar un vocabulario en forma de taxonomía. Resultados preliminares muestran una precisión de 71,57% en hiperónimos y de 67,97% en hipónimos.

Palabras clave: estadística de coocurrencia, extracción de taxonomías, lexicografía computacional, relaciones hiperonímicas

Abstract: We present a methodology based on co-occurrence statistics between headwords and words in their definitions in order to derive hypernymy relations from a lexicographic corpus, as part of a more extensive project devoted to the creation of a general purpose Spanish ontology of nouns and its application to the study of predicate-argument structures. The idea of the present proposal is to extract these semantic relations using a statistical technique that allows to combine diverse lexicographic resources. We find that hypernyms of a word are frequently used in its definitions and, similarly, its hyponyms usually are those which have frequent mentions to this word in their definitions. This creates a statistical association between words that allows for a taxonomic structuring of a vocabulary. Preliminary results show precision figures of of 71,57% in hypernyms and of 67,97% in hyponyms.

Keywords: co-occurrence statistics, computational lexicography, hypernymy relations, ontologías, taxonomy extraction

1 Introduction

This paper explores the possibility of using definiens-definiendum co-occurrence statistics to derive hypernymy relations from a lexicographic corpus. The idea of the present proposal is to try to extract hypernymy relations from a combination of diverse lexicographic resources using co-occurrence information and to propose a method to deal with problems related to polysemy, one of the most important challenges in ontology extraction methods.

It is a already an established idea that the meaning of a word can be deduced from the words related to it on the syntagmatic and paradigmatic axes (Harris, 1954; Sinclair, 2004, among others).

* This research was funded by project: "Agrupación semántica y relaciones lexicológicas en el diccionario", lead by J. DeCesaris (HUM2009-07588/FILO); APLE: "Procesos de actualización del léxico del español a partir de la prensa", 2010-2012, lead by M. T. Cabré (FFI2009-12188-C05-01/FILO). We would like to thank the anonymous reviewers for their comments and Chris Norrdin for proofreading.

ISSN 1135-5948

However, corpus based studies of predicate-argument relations must face the problem of retrieving the semantic information we cannot deduce directly from the context. Similarly as corpus linguists do when they lemmatize a corpus to obtain a more accurate count of the vocabulary frequency, in this type of distributional analysis of meaning we must classify the instantiated words in semantic classes. Consider, for example, the Spanish verb *calmar* 'to calm'. In the corpus, we find a large number of arguments in direct object position, such as *ansiedad, dolor, hambre, jóvenes, niños, sed, señora, temor*, 'anxiety, pain, hunger, young people, children, thirst, lady, fear', etc. It is easy to manually separate these nouns into two groups: a) human and b) feelings and appetites, and this is a necessary step for conducting the semantic analysis which will led us to conclude that the transitive structures of *calmar* are linked to two meanings: a) 'to make a person calm down' and b) 'to alleviate pain or necessity'. An ontology of nouns would be thus helpful for a faster and more accurate semantic analysis of corpus by replacing the arguments as they appear in corpus with their corresponding hypernyms (e.g., *a lady is a kind of human*, *anxiety is a kind of feeling*). It would be possible then to encode every particular instance of the verb occurring with each different kind of human or feeling to produce general patterns such as *to calm + human* or *to calm + feeling/appetite*. This operation would increase the power of generalization of corpus analysis by translating the myriads of possible arguments in more general and useful categories. This is why we need a wide coverage Spanish ontology or, more precisely, a device capable of producing hypernyms for any given input word instead of the static and limited ontological resources available such as the Spanish WordNet.

Using dictionaries to extract semantic information is a well-known methodology (Section 2), but the strategy of combining multiple lexicographic resources and treating them as a unified corpus, to our knowledge, is a novel approach. Using word frequencies in multiple dictionary definitions has been attempted before with the goal of extracting hypernyms (Nazar & Janssen, 2010). In this paper we elaborate on this simple idea and take it further to extract both hypernyms and hyponyms of a word by iteration of the algorithm that counts the frequency of co-occurrence between definiens and definiendum. In this approach, we reduce a group of dictionaries to a two-column matrix. One of the columns corresponds to headword entries and the other contains the vocabulary of all the definitions of such headword (including definitions for the different senses of the word as well as examples of use), in decreasing order of frequency. Thus, the interest lies in the study of the statistical association between words in both columns, without using linguistic or ontological knowledge and ignoring all other aspects of the definitions text such as syntactic parsing, which results in a very simple and yet robust and flexible mechanism to derive the taxonomy of a language on the fly.

Our experiment is part of a broader project devoted to the field of semantic analysis, one of the most important and challenging problems in NLP. The strategy presented here, as it is dictionary-driven and not corpus-driven, cannot be the only one used to conduct semantic analysis, but it can be part of a group of combined strategies. We are conducting extensive experiments on taxonomy induction using different methods of distributional semantics, of which the method described in the present paper is just a particular case. In Nazar & Renau (in press), we carry out an experiment with data from a corpus of general language text, and in Nazar & Renau (in preparation) we create clusters of words that have similar profiles of co-occurence and are, thus, semantically similar (in most cases sharing the same hypernym). As we will show in Section 5, the next step to follow is mixing these different experiments in order to improve the results. Finally, it is also important to emphasize that although our experiments are carried out in Spanish, there is nothing language specific in the approach and therefore it should be possible to replicate the experiment in other languages and with other lexicographic resources.

2 Related Work

Efforts in automatic taxonomy extraction have been reported for decades. First attempts involved the extraction of taxonomies from dictionaries, and were mainly based on the parsing of the definitions (Calzolari et al, 1973; Calzolari, 1977; Amsler, 1981;

Chodorow et al, 1985; Nakamura & Nagao, 1988; Wilks et al, 1989). These authors assume that, for instance, the first noun in the definition of a noun is in general the genus term, as in the example (1) for the definition of the English word *sedan* in the *Longman Dictionary of Contemporary English*, where the noun (*car*) will be treated as the hypernym of *sedan*:

(1) **Sedan**: a car that has four doors, seats for at least four people, and a trunk.

Of course, this reasoning will not be useful in slightly more complicated cases such as a definition of *truck* (2) in the same dictionary, where the first noun, "piece", cannot be considered the genus term:

(2) **Truck**: a simple piece of equipment on wheels used to move heavy objects.

According to Chodorow et al. (1985), these difficulties can be circumvented using what the authors call "empty heads", consisting of a closed list of non-content words that should be ignored (e.g., *piece of, variety of, type of*, and so on), which would result in the tuple *truck* IS-A *equipment.*

Exploiting dictionaries in this way has continued until the present (Gonçalo Oliveira et al, 2011). There is, however, another trend that started with the arrival of corpus linguistics, when authors started to explore the possibility of hypernymy extraction not from dictionaries but directly from corpora, applying lexico-syntactic patterns that explicitly convey hypernymy relations (Hearst, 1992). Thus, if a word W_1 is a hyponym of a word W_2, such patterns could be W_1 *is a kind of* W_2, W_1 *is a type of* W_2, W_1 *and other* W_2, among many others.

Strickingly, the vast majority of the work in taxonomy extraction has long disregarded the application of co-occurrence statistics or quantitative methods in general. Machine learning methods, which are quantitative by nature, have been applied for taxonomy extraction (Snow et al, 2006; Pantel & Pennacchiotti, 2006) but, again, only for the extraction of lexico-syntatic patterns *à la* Hearst. To our knowledge, no study takes the frequency of co-occurrence to increase the certainty of a given hyponym-hypernym pair. Naturally, if *sedan* appears multiple times in different instances of Hearst's patterns with the word *car*, then that should be taken as a clear indication the hypernymy relation is correct in that case.

Another aspect that authors in the field of taxonomy induction often ignore is the problem of polysemy, as pointed out by some authors (Guthrie et al, 1990; Klapaftis & Manandhar, 2010). If not addressed properly, polysemy can potentially harm the results of any taxonomy extraction attempt, because such taxonomy would fail to offer the inheritance and transitivity properties. Problems of polysemy are not limited to homographs. They can be of different natures, for instance, regular polysemy or systematic polysemy, among others (Pustejovsky, 1995; Agirre & Edmonds, 2006; Jezek & Hanks, 2010).

In parallel to advances in automatic taxonomy extraction, there are also relevant pieces of work related to manual development of taxonomies, as in the case of WordNet (Miller, 1995) and EuroWordNet (Vossen, 1998), probably the most widely known current hand-crafted project of this kind. In this tool, the general idea was to follow the criteria "is a" for connecting the hypernyms with its hyponyms. The reason why WordNet is not an adequate tool for our purposes is that it is based on the concept of "synsets" or "synonyms sets", which are groups of words connected by a common meaning. Every synset is connected to the others and all of them create the taxonomic tree. Thus, there is no lexical approach in this idea, but a kind of 'conceptual approach' not appropriate for the analysis of lexical units. If Spanish WordNet gives us a synset made from *delito, falta, fechoría, malhecho* 'crime, offense, misdeed, bad action' as equivalents of the English synset containing *misbehavior* and *misdeed*, it is based on a criterion than is too loose for lexical analysis, as lexical units as *delito* and *falta* cannot be considered exact synonyms. In other cases, the taxonomic hierarchy fails when selecting one specific lexical unit from the synset and trying to connect to other specific ones in the upper levels. In the case of *aguja* meaning 'stylus', we find that the English hypernym *device* has been translated into a synset with three hypernyms: *aparato, dispositivo, mecanismo*, but in Spanish this three words are very different from each other, and in the case of *aguja* 'stylus', it is probably closer to *dispositivo* and *mecanismo* than to *aparato.* All these problems derive from the synset-centric approach, which is not compatible with our lexico-centric approach. Authors such as Palmer (1998) and Hanks & Pustejovsky (2005) have raised similar objections.

Necessities of lexical analysis can be probably better achieved through methods such as Corpus Pattern Analysis (Hanks, 2004), which offers a systematic way of manually analyzing lexico-syntactic patterns. Patterns are constituted by syntactic information related to the arguments of the analyzed verb, and these arguments are also semantically analyzed by linking them to semantic types taken from an ontology. Thus, for example, in the case of the verb *calmar*, mentioned in Section 1, it would be divided into two different patterns as the following: [[Human 1]] calmar a [[Human 2]] and [[Human — Eventuality]] calmar [[Emotion — Appetite]]. This methodology is an important guide to our approach based on the Theory of Norms and Exploitations (Renau & Nazar, 2011), which offers a systematic way of manually analyzing lexico-syntactic patterns.

3 Methods

As stated in the introduction, our approach to the problem of taxonomy extraction is based on statistics of co- occurrence between defined words and the words in their definitions. The first step of our methodology is to compile a large lexicographic corpus from the web (Section 3.1). This corpus is then converted into a two-column matrix that encodes the frequency of co-occurrence of the defined words and those used in their definitions. With this matrix, we can compute two operations for any given input word: first, we look-up the word in the definiendum column of the matrix and then retrieve the most frequent words used in its definitions (Section 3.2). Second, we proceed exactly the other way round: the input word is looked up this time in the definitions of other words (Section 3.3). With the first operation we retrieve hypernym candidates, and with the second we retrieve hyponym candidates and also improve the certainty in the selection of both kinds of candidates by eliminating words that appear at the same time as hyponyms and hypernyms candidates. With respect to the problem of polysemy in the assignment of hypernymy (Section 3.4), our approach is based on an iteration of the same process: for every particular hypernym candidate proposed for a given input word, the system retrieves all the other probable hyponyms. Then, by representing these relations in the form of a directed graph, we observe a natural clustering of the words according to their different senses.

3.1 Compilation of a lexicographic corpus

Experiments were conducted on a lexicographic corpus crawled from the web. With this corpus, we create an index that registers words that tend to appear in the definitions of other words. With this index, we derive conclusions such as the hypernymy structure of a vocabulary not from a single lexicographic authority but from the aggregation of a multiplicity of sources. This makes up for the fact that the downloaded corpus is rather noisy and that many of the definitions are not entirely satisfactory if taken in isolation. The aggregated material as a whole, in contrast, offers better certainty as a consequence of the cumulative effect of the definiens-definiendum co-occurrence.

3.2 The input word in the definienda

In this phase of the procedure, the analysis of co-occurrence of words means to obtain, for any given word, a list of the most frequent content words that occur in its definitions. Thus, for instance, in the case of the word *sedán* 'sedan' the most frequent word in all the definitions is *automóvil* 'car'. Similarly, the most frequent word in the definitions of *triquinosis* 'trichinosis' is *enfermedad* 'disease'. This is a very stable pattern, however it is not sufficient for the development of a full scale ontology, and that is why we still need to carry out further operations.

3.3 The input word in the definiens

In this phase, we take the result of the previous one and iterate it. For each hypernym candidate obtained, we analyze in which other definitions they appear (definitions of words other than the initial input word). This gives us the possibility to find not only hypernym candidates (which we represent by outgoing arrows) but also hyponym candidates (represented by incoming arrows). To use one of the previous examples, Figure 1 shows a graph in which the initial input word *sedán* is linked to the hypernym candidate *automóvil* 'car'. This hypernym, in turn, has links (incoming arrows) from different words apart from *sedán.* These are other words that, similarly as *sedán*, also have *automóvil* 'car' as the most frequent word in the definitions, e.g. *taxi, camión, berlina, grúa* 'taxi, truck, berlin, wrecker'. Of course, not all of these links are correct. There are also elements which are parts of an automobile and are confused with hyponyms: *luz, luneta, intermitente* 'light, rear-view window, indicator light', etc. In the case of *sedán* itself, unsurprisingly it does not appear to have hyponyms (it is not the most frequent word used in the definitions of any other word). Finally, the graph also reflects the relationship between *automóvil* and a higher order hypernym *vehículo* 'vehicle', thus, a hierarchy of two levels. This new link is derived from *automóvil* by an iteration of the same process that started from *sedán*, i.e., *vehículo* 'vehicle' is the most frequent word in the definitions of *automóvil.*

We could continue to iterate the process, for instance with *vehículo*, but that would depend on the particular task at hand. Thus, the number of iterations of the process is an execution parameter.

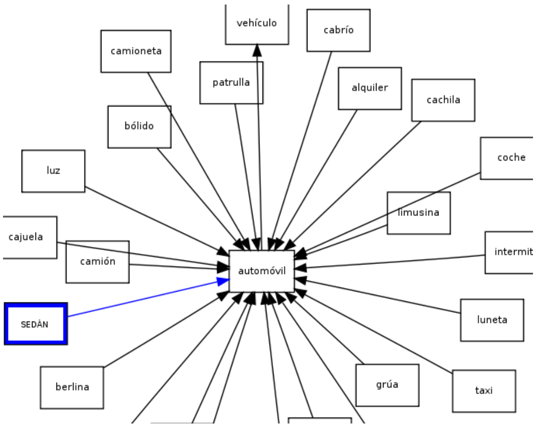

Figure 1: The noun *sedan* linked to *automóvil* as its hyponym and to other kinds of vehicles as co-hyponyms. At the same time, *automóvil* is linked to *vehículo* as its hyponym.

3.4 Word sense induction

Our attempt to solve the problems of polysemy by the use of word co-occurrence graphs is motivated by a distinctive geometric property of these graphs, which is to have attractors or hubs, defined as regions of the graph depicting nodes with large numbers of incoming arrows. In the case of word co-occurrence graphs, these hubs naturally represent the different senses of a word and can be of help in the process of disambiguation.

As we can see in Figure 2, for the case of a polysemous word such as *langosta* ('locust'/'lobster'), there are certain nouns in the network that have an important number of incoming links. Notice that different nouns co-occurring with them are clustered. These clusters represent the two different meanings of the word *langosta*, one for the land animal (*locust*) and the other one for the marine animal (*lobster*). Both groups are created around the hypernyms *insecto* ('insect') and *crustáceo* ('crustacean').

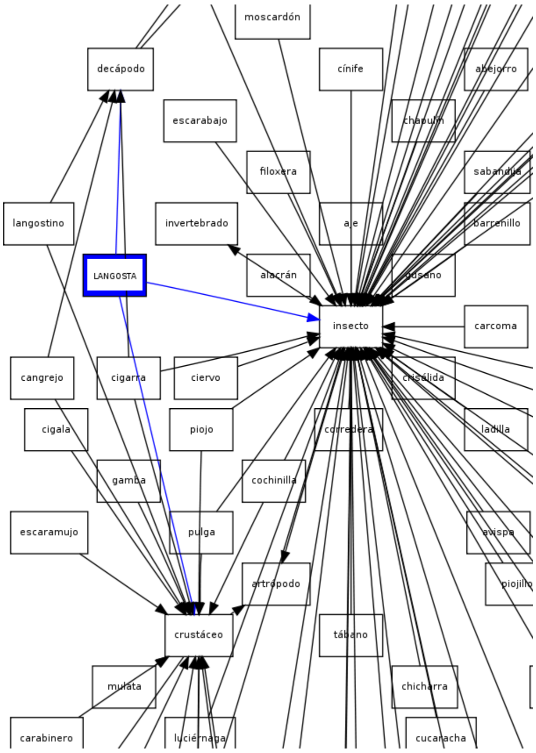

Figure 2: Fragment of the co-occurrence graph of the noun *langosta*, divided into two general meanings (clusters): *insecto* in the 'locust' sense and *decápodo* or *crustáceo* in the 'lobster' sense.

4 Results and Evaluation

For the evaluation of the system, we manually analyzed 173 nouns from the taxonomy, divided in the following groups:

a) 8 nouns typically used as hypernyms of many other nouns, in order to know how hyponyms were detected by the system: *calzado* 'footwear', *mueble* 'furniture', *embarcación* 'vessel', *queso* 'cheese', *herramienta* 'tool', *sombrero* 'hat', *mineral* 'mineral', *utensilio* 'utensil'.

b) 15 polysemous nouns with at least two clear different meanings, in order to get information about how the hypernyms were detected (the English translation is given only for the most typical meaning): *águila* 'eagle', *manta* 'blanket', *aguja* 'needle', *mono* 'monkey', *araña* 'spider', *ratón* 'mouse', *dragón* 'dragon', *rémora* 'remora', *emperador* 'emperor', *sirena* 'siren', *fraile* 'monk', *tritón* 'newt', *gato* 'cat', *zapatero* 'cobbler', *langosta* 'lobster'.

For example, in the case of the word *águila*, we would expect to find the meanings 'eagle' and 'sharp person', being thus the hypernyms *ave* 'bird' and *persona* 'person', respectively. Similarly, in the case of *sirena* 'siren', we should see at least the meanings of 'piece of equipment' and 'aquatic nymph', etc.

c) 150 nouns randomly sampled from the dictionaries.

Despite the fact that groups *a* and *b* were created to focus on the problem of hypernymy and homonymy separately, both semantic relationships were evaluated in the three groups. The basic criterion used to evaluate the results is that nouns linked to a hypernymy relationship must strictly accept the test "is a" or "is a kind of". Thus, for instance in the case of *calzado* 'footwear', it could be possible to create a sentence like "A *bota* 'boot' is a kind of *calzado*", in which *bota* is a hyponym candidate.

The results of the overall evaluation are summarized in Table 1. The algorithm detected correctly 71.57% of the hypernyms and 67.97% of the hyponyms. For 42 nouns (24%) there were no results.

	Correct		Incorrect		Total
	n	%	n	%	
Hypernyms	214	71.57	85	28.43	299
Hyponyms	399	67.97	188	32.03	587
					886
No results:	42 (24 %)				

Table 1: Results of hypernyms and/or hyponyms candidates.

For group *a*, the algorithm detects many types of objects related to the hypernyms evaluated, that is, it detects *brie, cabrales, camembert, feta*, etc., as kinds of cheeses; *abanico, cuchara, grapadora, matamoscas* 'fan, spoon, stapler, flyswatter', etc., as kinds of utensils, and so on. In the case of group *b*, of polysemous words, co-occurrence graphs detect the most prototypical meaning in the majority of the cases. Table 2 shows the meanings detected for every polysemous word taken from the evaluation.

Nouns	**Hypernyms detected**
águila	2: *ave, moneda*
aguja	4: *barra, instrumento, pez, varilla*
araña	4: *arácnido, lámpara, planta, red*
dragón	6: *animal, embarcación, pez, planta, reptil, soldado*
emperador	4: *dignidad, pez, soberano, título*
fraile	2: *monje, montón (de uva)*
gato	2: *mamífero, persona*
langosta	3: *crustáceo, decápodo, insecto*
manta	2: *pieza, tela*
mono	1: *persona*
ratón	2: *dispositivo, mamífero*
rémora	1: *pez*
sirena	3: *aparato, instrumento, ninfa*
tritón	2: *anfibio, dios*
zapatero	4: *insecto, mueble, persona, pez*

Table 2: Number of hypernyms detected in 15 of the candidates (group *b*).

In order to estimate recall, we measured the coincidence of hypernyms and hyponyms of groups *a* and *b* with those provided by WordNet. To obtain this estimation we first had to manually check if the units provided by WordNet were indeed correct, and only then we compared those with the result of our algorithm. In total, WordNet provides 211 units (only in the first level and leaving aside multiword terms), but 49 of them are incorrect (e.g., we find nouns such as *diapasón* 'diapason', *remo* 'paddle' or *cepillo* 'brush' as hyponyms of *herramienta*, which is not exact because *herramienta* is translated from 'implement', but 'implement' can have other meanings as well, such as *utensilio*, probably closer to the mentioned hyponym candidates). For the remaining 162 units from WordNet, 55 are also in our taxonomy (34% recall). This figure, however, is only approximate because there are also units in our taxonomy that are correct and are not included in WordNet.

With respect to error analysis, the most important problems derive from the confusion between semantic relations: a) confusion hypernym-hyponym, e.g. *ratón* 'mouse' is taken as the hypernym of *múrido* 'murine', when in reality a mouse is a kind of murine and not vice-versa; b) confusion hypernym/hyponym-synonym, especially in the case of slang variants, e.g. *minino* 'kitty' is put in the place of the hyponym of *gato* 'cat', when it is actually a colloquial synonym; confusion hypernym/hyponym-meronym/ holonym, e.g. *oro* 'gold' is taken by the hypernym of *águila* because the latter can be a kind of coin made of gold, thus the material of the coin is taken as if it was a hypernym. In other cases, the reason for the incorrect results are basically due to the lexicographic nature of the data, in which abbreviations, etymology or other informations are offered. Sometimes, the absence of data represents the lack of consensus in the different dictionaries, e.g. in the case of *rémora* taken as a 'hindrance', definitions were varied, and the concept was defined with hypernyms such as *cosa* 'thing', *obstáculo* 'obstacle' or *impedimento* 'impediment'.

5 Conclusions and Future Work

This paper has shown a language independent statistical method which uses a large number of online dictionaries as input to create a corpus-driven ontology with no human supervision. The results shown in the previous section give an approximate idea of the usefulness and limitations of the system. On the one hand, a 67-71 of precision seems to be sufficient as a basis of a ready-to-use tool to complement and facilitate the analysis by human expert. At the same time, the recall obtained seems to be appropriate for detecting the main meanings of a noun, that is, the meanings in which different lexicographical authorities agree upon, and in this sense the algorithm can be used as a reference for basic semantic analysis. On the other hand, many important meanings of some words were not detected, and often it is in these cases where the lexical analysis is most needed, and not in the main or more frequent meanings.

A general limitation of the strategy has already been pointed out in Section 1: as the sources used to run the experiment are dictionaries, the strategy lacks a real corpus-driven approach with direct contact with real data. In this sense, we are confident that the strategy presented in this paper can be combined with other strategies in order to create a ready-to-use taxonomy, having the advantage, in comparison with other dictionary-based methodologies, of representing the consensus and variety of information founded in diverse lexicographical resources.

Among the lines of future work, extensive evaluation of our word disambiguation strategy has to be carried out. We will also address the study of multiword expressions, which were not included in this paper. Another remaining factor is that our solution proposes for each word different hypernym candidates at each link of the chain, as in the example of *langosta* described in Section 3.4, where we had different hypernyms according to the 'locust' or 'lobster' senses. Strictly speaking, this is not a solution for the structural problem that polysemy represents for an ontology. In this taxonomy, we could start from a given word such as *tarántula* 'tarantula' and go up one level to a correct hypernym, such as *araña* 'spider', and then to an upper level hypernym such as *arácnido* 'arachnid'. However, being *araña* a polysemous word, we should not also end up in an incorrect hypernym such as *lámpara* (see Table 2). In a semasiological approach like ours, the solution for the problem of polysemy has to be tackled differently in every particular task at hand, that is, it has to be treated as a problem of disambiguation. This can include the computation of distributional similarity coefficients between the context of a particular instance of a word and those contexts of the different senses that the word can have. Again with the *langosta* example, our ontology should be able to assign a correct hypernym for a target word depending on the words that are found in the context, depending on whether they are related to insects or to lobsters.

References

Agirre, E. & Edmonds, P. (eds.) 2006. Word Sense Disambiguation. Algorithms and Applications. Dordrecht, Springer.

Amsler, R. 1981. A taxonomy for English nouns and verbs. Proc. of 19th annual meeting on ACL (Morristown, NJ, USA): 133–138.

Calzolari, N. 1977. An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie* 31(2): 118–128.

Calzolari, N., Pecchia, L. & Zampolli, A. 1973. Working on the Italian machine dictionary: a semantic approach. Proc. of 5th Conference on Computational Linguistics (Morristown, NJ, USA): 49–52.

Chodorow, M., Byrd, R. & Heidorn, G. 1985. Extracting semantic hierarchies from a large on-line dictionary. Proc. of the 23rd annual meeting on ACL (Chicago, Illinois, USA): 299–304.

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

Gonçalo Oliveira, H., Antón Pérez, L., Costa, H. & Gomes, P. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* 3(2): 23–38.

Hanks, P. 2004. The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography* 17(3): 245–274.

Hanks, P. & Pustejovsky, J. 2005. A Pattern Dictionary for Natural Language Processsing. *Revue française de linguistique appliquée* 10(2): 63–82.

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

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. Proc. of the 14th International Conference on Computational Linguistics (Nantes, France): 539–545.

Jezek, E. & Hanks, P. 2010. What lexical sets tell us about conceptual categories. *Lexis* 4: 7–22.

Klapaftis, I. & Manandhar, S. 2010. Taxonomy Learning Using Word Sense Induction. Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (Los Angeles, USA).

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

Nakamura, J. & Nagao, M. 1988. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. Proc. of the 12th International Conference on Computational Linguistics COLING-88 (Budapest, Hungary): 459-464.

Nazar, R. & Janssen, M. 2010. Combining Resources: Taxonomy Extraction from Multiple Dictionaries Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10): 1055–1061.

Nazar, R. & Renau, I. In press. A Co-occurrence Taxonomy from a General Language Corpus Proc. of EURALEX 2012. (Oslo, 7-11 August, 2012).

Nazar, R. & Renau, I. In preparation. Agrupación semántica de sustantivos basada en similitud distribucional. Implicaciones lexicográficas. V Congreso Internacional de Lexicografía Hispánica (Madrid, 25-27 June, 2012).

Palmer, M. 1998. Are WordNet sense distinctions appropriate for computational lexicons? SIGLEX-98, SENSEVAL (Herstmonceux, Sussex, UK, Sep 2-4, 1998).

Pantel, P. & Pennacchiotti, M. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. Proc. of 21st International Conference on Computational Linguistics and 44th annual meeting of the ACL (Sydney, Australia): 113–120.

Pustejovsky, J. 1995. The Generative Lexicon. Cambridge: MIT Press.

Renau, I.; Nazar, R. 2011. Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis. Proc. of the 27th Conference of SEPLN (Huelva, September 5-7, 2011).

Sinclair, J. 2004. Trust the Text: Language, Corpus and Discourse Routledge.

Snow, R., Jurafsky, D. & Ng, A. 2006. Semantic taxonomy induction from heterogeneous evidence. Proc. of the 21st International Conference on Computational Linguistics (Sydney, Australia): 801–808.

Vossen, P. 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. *Computers and the Humanities* 32(2-3).

Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T., Slator, B. 1989. A Tractable Machine Dictionary as a Resource for Computational Semantics. Computational Lexiography for Natural Language Processing. B. Boguraev and T. Briscoe (eds): 193-228. Essex, UK: Longman.