A co-occurrence taxonomy from a general language corpus[1]

Rogelio Nazar & Irene Renau

Keywords: *asymmetric word association, computational lexicography, co-occurrence statistics, distributional semantics, taxonomy extraction*.

Abstract

This paper presents a quantitative approach to the generation of a taxonomy of general language. The methodology is based on statistics of word co-occurrence and it exploits the fact that word association is asymmetrical in nature, in much the same way as hyperonymy relations are. Words tend to be syntagmatically associated with their hyperonyms, though this is not true the other way round. Taking advantage of this phenomenon, and with the help of directed graphs of word co-occurrence, we were able to collect hyperonym-hyponym pairs using a reference corpus of general language as the only source of information, i.e., without using lexico-syntactic patterns nor any kind of pre-existing semantic resources such as dictionaries, ontologies or thesauri. The results obtained by using this method are not precise enough to be used for immediate practical purposes, but they confirm the hypothesis that as a general rule hyperonymy is linked to asymmetric co-occurrence relations. The paper discusses an experiment in Spanish, but we believe the same conclusions apply to other languages as well.

1. Introduction

Among all the important aspects of the work of a lexicographer, one crucial task is that of conducting semantic analyses. Today, when the corpus-driven approach is widely established in the discipline, the lexicographer has an array of corpus exploitation tools at his or her disposal and can apply them to obtain data about the real use of words. Typically, such an analysis consists in the extraction of a variable number of concordances in order to obtain semantic information related to the definiendum. This procedure is labor-intensive and time-consuming and, moreover, it entails the risk of not being sufficiently systematic.

In this paper, we explore the possibility of providing lexicographers with information derived from the corpus about the hyperonymy of nouns. It goes without saying that an organization of the vocabulary into hyperonym-hyponym pairs would be valuable initial general information and would save considerable time and effort in the early phases of a lexicographic project. Furthermore, hyponyms linked to a noun are useful to create coherent groups of words that could be written with the same pattern of definition (for instance, varieties of cheese, mammals, hats, and so on). Finally, the proposed method could also be useful for other projects, such as the organization of a textual corpus into an ontology.

Our proposal is based on the idea that word co-occurrence is asymmetric in the case of hyperonym-hyponym pairs. That is to say, a given word might show a tendency to co-occur with its hyperonym, but this is not true the other way round. For instance, it is more probable that words such as *motorbike, bicycle* or *truck* will co-occur with their hyperonym *vehicle* than *vehicle* to co-occur with any of its particular hyponyms. Using word co-occurrence graphs, we have been able to exploit these asymmetric relations and the result has been a hierarchical organization of the lexicon emerging naturally from the corpus. It should be said, however, that the most important aspect of the present proposal is not the resulting taxonomy per se, but rather the method used to generate it, because in this early stage of our research we are not presenting a finished product but rather testing an idea that we think has wider implications. Above all, the taxonomy we present is interesting because it is a rather complex elaboration which started from very simple input, such as the fact that two lexical units have a tendency to co-occur in the same sentences. We would like to test how far we can advance

with this kind of elementary data before starting to add language-specific code into the system. As long as our approach is not language specific, it should be possible to obtain the same result with a replication of the same experiment in other languages. In this article we report on experiments evaluating this particular co-occurrence approach in a general Spanish corpus. A finished Spanish taxonomy, which is left for future work, would have to be the result of a combination of different techniques.

2. Related work

The first attempts to derive a taxonomy automatically were based on the use of machine readable dictionaries, and the strategy was to study lexical patterns which express hyperonymy relations within the definitions (Chodorow et al. 1985, Alshawi 1989, among many others). These studies inspired further research into the extraction from dictionaries of not only hyperonyms but also a variety of other semantic relations (see, for instance, Fox et al. 1988 and Dolan et al. 1993).

In the nineteen nineties, the interest in taxonomy extraction shifted from dictionary-based to corpus-driven approaches, when authors started to use hand-crafted lexico-syntactic patterns to extract hyperonymy relations from corpora (Hearst 1992, Rydin 2002). This other method essentially consists in looking in corpora for occurrences of patterns such as ***X*** *is a kind of* ***Y***, assuming then that any pair of nouns occupying the positions of ***X*** and ***Y*** will hold a hyperonym-hyponym relationship. Snow et al. (2006) proposed a variant of this approach, using machine learning to derive these lexico-syntactic patterns from the corpus automatically by means of pairs of hyperonym-hyponym seed terms.

A totally different approach has been laid out by authors who prefer to use the term *thesauri* instead *taxonomies* (Grefenstette 1994, Schütze & Pedersen 1997, Lin 1998, and others). In contrast to the above-mentioned studies, they use a method which is quantitative in nature. Here, the rationale is not to find hyperonym-hyponym pairs, but instead to find groups of words which are distributionally similar and, therefore, liable to be placed in the same category (under the same hyperonym). If it is possible to group similar words into a same set, then it suffices to find a correct hyperonym for at least one member of the group in order to find the right hyperonym for the whole category.

In parallel to this research in computational linguistics, other efforts have been taking place for the manual compilation of large taxonomies, the most widely known being projects such as WordNet (Miller 1995, Fellbaum 1998) and EuroWordNet (Vossen 1998), which link hyperonyms, hyponyms and other semantic relationships, as well as offering definitions and joining words in synonym sets.

The present paper differs from previous attempts reported in the literature. Our motivations are different to those underlying the manual development of taxonomies, since we are interested in deriving the taxonomy not only automatically but, most importantly, from the corpus, i.e., as the result of the text produced by a linguistic community and not mediated by a particular individual. We use the term ‘natural taxonomies’ to designate this kind of co-occurrence network. A large body of work has been produced on the subject of lexical co-occurrence, especially after the work of Church and Hanks (1991). However, and probably as a consequence of the fact that these last two authors used a symmetric association measure like mutual information, the asymmetry of word co-occurrence – which was already noticed in earlier research (Phillips 1985) – remains unexplored. In particular, to our knowledge there is no precedent on the use of the property of asymmetry to derive taxonomies in the way outlined in this proposal. This approach is also different from those based on lexico-syntactic patterns, applied to both machine-readable dictionaries and corpora,

and in the same way it cannot be compared to distributional thesauri methodologies, since we are not extracting similar words but rather hyperonym-hyponym pairs. In relation to dictionary-based methods or to the models based on lexico-syntactic patterns, the approach presented here is clearly corpus-driven and it is not dependent on previous linguistic analysis, like dictionaries, or the already-elaborated definitions one can find in corpora. A similar approach was used before (Nazar 2010; Nazar et al., forthcoming) to derive taxonomies for specialized terminology in English, but a general language taxonomy represents a far more difficult scenario. In technical and scientific literature, concepts are more rigorously defined, whereas ordinary language has a variety of purposes other than that of classifying concepts into more general categories, and thus the corpus is expected to contain less useful information.

Finally, with regard to related work, a previous study (Renau and Nazar 2011) explains the wider context of this research, consisting in the analysis of lexico-syntactic patterns derived from corpora. The motivation was to extract the meaning/s of words (verbs, above all) as they are used in real contexts, both for lexicographical purposes and for pure semantic analysis. In that study, in a line similar to the work of Firth (1957), Halliday (Kress 1976), Sinclair (2004) and Hanks (2004), we found that we could improve semantic analysis by replacing the arguments of the verbs by their hyperonyms, a move that increases the power of generalization of the corpus. Thus, for instance, statistics of co-occurrence inform that *calarse el sombrero* ('to pull down one's hat') is a normal verb-noun combination in Spanish. However, if we can provide the system with the ability to detect hyperonyms of the nouns, then the system can recognize that words such as *sombrero, gorra, boina,* etc., ('hat, cap, beret', etc.) are the same type of object and can appear on similar contexts (it is also possible to say *calarse la gorra* or *calarse la boina*).

3. Methods

As stated in the introduction, our approach to the problem of taxonomy extraction is based on statistics of word co-occurrence. The experiment we describe needs two types of input data to be undertaken: 1) a list of words to be placed in a taxonomy, and 2) a reference corpus of general language. It is possible to derive 1 from 2, yet for the purpose of this experiment we used different arbitrarily selected lists of nouns. With respect to 2, we used a sample of approximately 120 million words of general Spanish text, consisting of a random selection of articles from Wikipedia. This corpus represents about 25% percent of all the text that comprised Wikipedia in the year 2010, all metadata information related to the structure and categories of the pages as well as html links and the rest of the internal code being eliminated, leaving only a plain string of characters from the body of the texts. The motivation behind this procedure is that we are not interested in Wikipedia in particular – we only use it as a source of text in general. It is true that encyclopedic text is not general language, but as a large random sample of text we can assure that it is not text about a particular subject. We are aware of the fact that, because of its nature, the corpus contains much information about the elements described in the texts, but further experimentation with other genres (such as essays, press articles and fiction) will have to be reserved for future work.

With these materials at hand, we conducted a three-phase procedure, as follows. For every word in the input list, we first conduct an analysis of the co-occurrence of the word. This means obtaining words which are syntagmatically associated with the input word; in practice, this refers to a list of the most frequent content words that occur in the corpus in the context of the input word (three sentences, in our case). In order to consider only nouns, and

to account for their inflectional variation, the contexts of occurrence are lemmatized and POS-tagged using Schmid's (1994) Tree-tagger. A further refinement of this list is undertaken by eliminating words that are frequent but not statistically significant, which can be performed using the same reference corpus. Assuming that $f_o(i)$ is the observed frequency of word *i* in the context of a target word, and that $f_e(i)$ is the total frequency of the same unit in the whole corpus, the process of filtering is to eliminate all units with a w(*i*) score below an empirically determined threshold (in this case, -11), defined in the following equation:

$$w(i) = \log(\frac{f_o(i)}{(f_e(i)+1)})$$

Just to illustrate the point, consider an example with the input word *ciclomotor* ('moped'). The first step is to extract contexts of occurrence of this word in the corpus and to sort the vocabulary in those contexts in decreasing order of frequency, eliminating uninformative words with the aid of the above-mentioned coefficient. In the case of *ciclomotor*, there are only 13 contexts of occurrence in our corpus. The procedure eliminates common co-occurring words such as *año* ('year'), *tiempos* ('times'), *países* ('years'), but other content words as well, such as *motor* ('motor') and *modelo* ('model'). The remaining units are proper nouns such as *Betavus*, *Ducati*, *Yamaha* and *Vmax* but also words such as *bicicleta* ('bicycle'), *motocicleta* ('motorbike') and *década* ('decade'). The correct hyperonym (vehicle) is not shown in this group, which is not surprising in a sample of just a few sentences. However, it is thanks to these syntagmatically related words that we will be able to find it, since we will inspect the second-order co-occurrence of *ciclomotor* by observing which words tend to appear in the contexts of the words that appear near *ciclomotor*.

The analysis of second-order co-occurrence is similar to the previous, but it reveals that *vehículo* ('vehicle') can be frequently found in the contexts of some of these words, such as *motocicleta* and *bicicleta*. In 500 sentences from our corpus, the word *bicicleta* co-occurs most frequently with words such as *montaña* ('mountain'), *ciclismo* ('cycling'), *vasco* ('Basque'), *rueda* ('wheel'), *Hoffman*, *ciclista* ('cyclist'), *ruta* ('road'), *vehículo* ('vehicle'), etc. The word *motocicleta*, in turn, occurs in 302 sentences, and frequently co-occurs with *Honda*, *vehículo*, *Lizzie*, *Ducati*, etc. Therefore, we conclude that *ciclomotor* has a significant second-order co-occurrence with *vehículo*, and we can calculate the significance of this association using the wD score, defined as follows, where D(*i*,*j*) accounts for the number of times the word *j* was found in the lists of words co-occurring with the words that co-occur with the target word *i*, while $f_o(i,j)$ denotes the total frequency of *j* in the lists of *i*.

$$\mathrm{wD}(i,j) = \log(1 + f_o(i,j) * D(i,j))$$

Essentially, what this score does is to promote those lexical units which not only co-occur frequently with the target word, but are also well distributed within all the frequency lists that are generated from the target word.

The final step is to draw the taxonomy graphs using the information generated in the previous phases. The result of the first two phases is a matrix *M* that contains information about words that have a tendency to co-occur with other words, and what we do now is to represent that matrix as a directed graph. Thus, for a given word *A*, we start by drawing a node for this word, and from there we draw edges to the nodes of other words, indicating that *A* has a tendency to co-occur with these, according to matrix *M*. As a consequence of the fact that edges are directional, we observe that, as a general rule, the word that ends with more incoming arrows is the hyperonym of the target word and, when it is not, it results in a semantically related concept, such as a co-hyponym, meronym, synonym, or others. In Figure 1, hyperonym-hyponym relationships are indicated by the arrow connecting the word at the

top (hyponym) with the word at the bottom (hyperonym). For the purpose of our experiment, for any input word we only consider as output the node with more incoming arrows, ignoring the rest. Therefore, such graph is not meant to be interpreted as if *bicicleta* (‘bicycle’) were considered the hyperonym of *motocicleta* (‘motorbike’). The graph only shows that: 1) *ciclomotor* has a significant frequency of (first- or second-order) co-occurrence with *motocicleta*, *bicicleta* and *vehículo*; 2) *motocicleta* displays the same pattern with respect to *bicicleta* and *vehículo*; 3) *bicicleta* is also associated with *vehículo*; and, finally, 4) that *vehículo* does not show any tendency to co-occur with *ciclomotor*, *motocicleta* or *bicicleta*. In case different candidates have the same score, then the algorithm will select as hyperonym the one that has been selected more times by other words of the input list. This is important because it means that different results will be obtained as more input words are used in the experiment.

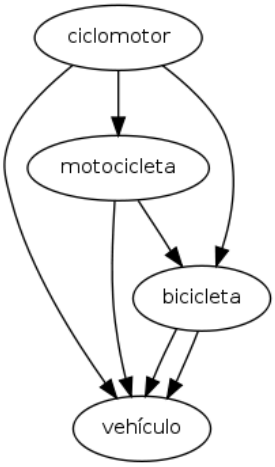

Figure 1. Example of a directed graph representing asymmetric lexical co-occurrence for the word *ciclomotor* (‘moped’) with respect to its hyperonym *vehículo* (‘vehicle’).

4. Evaluation

In order to evaluate our strategy, we used a sample of 200 arbitrarily selected Spanish words related to the semantic groups of mammals, insects, drinks, hats, vehicles and varieties of cheese. We submitted this list to the process explained in the previous section. For this first experiment, we selected nouns that are always part of a typical general dictionary because we started out from the premise that the list of headwords would be already selected.

The criterion for considering an outcome as correct or incorrect is simply to test whether a valid taxonomic relation holds between each input word and the proposed hyperonym. Thus, for instance, graphs reflecting the following relationships were considered correct:

brie (‘brie’) → *queso* (‘cheese’)
camión (‘truck’) → *vehículo* (‘vehicle’)
avispa (‘wasp’) → *insecto* (‘insect’)
agua (‘water’) → *bebida* (‘drink’)

We also considered correct connections to be those with less rigorously defined semantic classes in cases where a hyponym is linked to collective nouns, such as in the following cases:

castor (‘beaver’) → *fauna* (‘fauna’)
chocolate (‘chocolate’) → *gastronomía* (‘gastronomy’)

pulga (‘flea’) → *biología* (‘biology’)

In these cases, we expected to link *castor* with *mamífero* (‘mammal’), *chocolate* with *bebida* (‘drink’) and *pulga* with *insecto* (‘insect’), but the system suggested wider categories, comparable to those occupying the highest positions in a taxonomy. Thus, for example, *fauna* is used as a label in Spanish Wordnet, *gastronomía* could be easily assimilated to the label *alimento* (‘food’) and *biología* to *ser vivo* (‘living being’).

We repeated the experiment 4 times, adding 50 words to the input list each time in order to see how performance changed as more elements are added to the experiment, the results being those shown in Figure 2 and Table 1. As is common in information retrieval, in this experiment precision is defined as the number of correct cases over their sum with the incorrect cases, while recall is defined as correct cases over all cases (correct + incorrect + trials without answer).

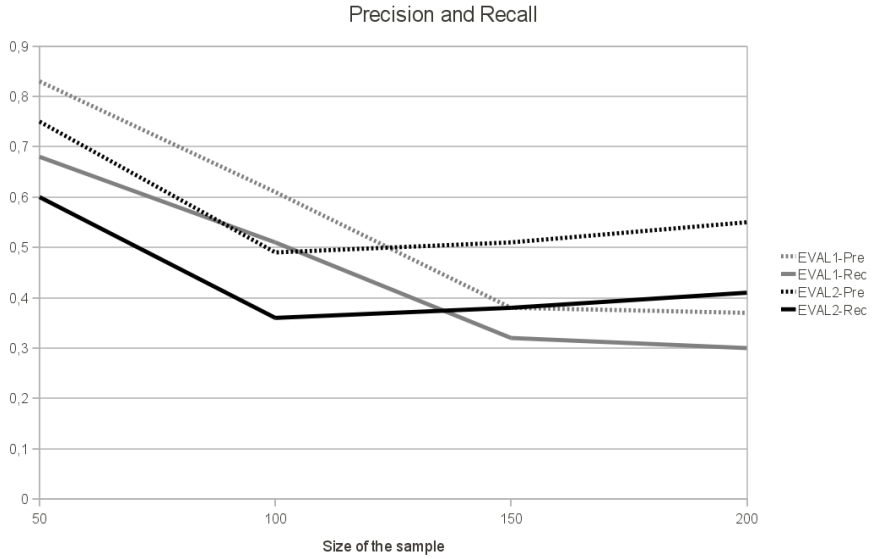

Figure 2. Precision and recall score for input samples of different sizes.

Table 1. Results of two experiments for evaluation.

Evaluation 1 (max. 250 contexts per word)				
Number of input words	**50**	**100**	**150**	**200**
Correct	34	51	48	60
Incorrect	7	33	78	104
Without answer	9	16	24	36
Precision	0.83	0.61	0.38	0.37
Recall	0.68	0.51	0.32	0.30
Evaluation 2 (max. 500 contexts per word)				
Number of input words	**50**	**100**	**150**	**200**
Correct	30	36	57	81
Incorrect	10	37	54	66
Without answer	10	27	39	53
Precision	0.75	0.49	0.51	0.55
Recall	0.60	0.36	0.38	0.41

The only difference between Evaluation 1 and Evaluation 2 is that in the first case we restricted the experiment to the extraction of a maximum of 250 contexts of occurrence per input word, while in Evaluation 2 we increased the maximum number of contexts to 500. It can be seen that in both cases the best results are obtained when there are only 50 words to classify (with 0.75-0.83 precision and 0.60-0.68 recall). An important number of incorrect results are obtained when the sample of words to be classified is increased to 100. In the case of Evaluation 2, from that point on the proportion of correct cases increases as more words are added to the sample. Future work will determine whether performance increases as more input words and more contexts of occurrence per input word are added to the experiment.

Among the most frequent types of errors there are those attributable to the effect of polysemy. For instance, for the word *caravana* ('caravan') for which we expected the hyperonym *vehículo* ('vehicle'), the system proposed instead the hyperonym *expedición* ('expedition, convoy'), which is a correct hyperonym but for a different sense of the word. Similarly, other errors were produced by homonymy, such as the case of the word *ron* ('rum', in the group of drinks), which was linked to the hyperonym *personaje* ('character') due to the proper noun *Ron Weasley*, the famous co-protagonist of *Harry Potter*. The majority of the incorrect hyperonyms, however, were part of the same semantic group as the hyponyms they were linked to. For instance, the system frequently connected mammal-hyponyms with mammal-hyperonyms, insects with insects, and so on. Thus, for example, in one of the experiments *camión* ('truck') was connected to the wrong hyperonym *tranvía* ('tram'), although both of them are 'vehicles'; *coyote* ('coyote') was linked to *venado* ('deer'), both being 'mammals'; and so forth. Another type of error is related to the features of the hyponyms, the most of the cases belonging to the 'formal quale', to use Pustejovsky's (1995) terminology. Thus, some hyponyms in the 'alcoholic drink' category were connected to *azúcar* ('sugar'), one of the common ingredients of many alcoholic drinks (we can consider it as a meronymy relationship as well); *chistera* ('top hat') was wrongly linked to *negra* (the feminine for 'black'), as this is the most common color for this kind of hat; and so forth. Finally, a small group of wrong hyperonyms had no relationship with the hyponym: *pamela* ('lady's wide-brimmed hat') was joined to *volumen* ('volume'), *sidra* ('cider') to *queso* ('cheese'), *reno* ('reindeer') to *oeste* ('west'), and so on.

With regard to figures of performance, it is also important to stress the fact that they are perhaps excessively conservative, because in many cases we are counting as incorrect results cases that in a real application could be considered useful. For instance, in the second evaluation, close to 15% of the errors were made when different animals where classified as hyponyms of *venado* ('deer') or *jabalí* ('wild boar'), which were in turn classified as animals. We have to consider them incorrect results in order to be consistent with the evaluation, but in a practical application what would be important is that animals are placed correctly in the same category, and this is precisely what we obtain. In practice, the fact that there is an incorrect link (*venado* or *jabalí*) between them and the correct hyperonym would be an issue of secondary importance.

5. Conclusions and future work

This paper has outlined a method for extracting hyperonym-hyponym pairs from textual corpora using only statistics of word co-occurrence and excluding all semantic resources other than lemmatization and POS-tagging. Even when the results are not sufficiently accurate to be used as a finished product, they open the way for a new approach to the problem of taxonomy extraction. Further refinements of the strategy may improve the quality of the results. In any case, a partially correct taxonomy can be used as a starting point for its

development by lexicographers, as less effort is required to correct information than to generate it.

There is a double motivation behind this project, one theoretical and one application-oriented. From the theoretical point of view, we follow and attempt to extend Harris's (1954) hypothesis on the relationship between distributional and semantic similarity. We have shown that, thanks to this idea, we can now have a semantic hierarchy naturally emerging from corpora. From a practical point of view, it is obvious that dictionaries and other lexical tools can benefit from the statistical treatment of data, both to improve and to accelerate the analysis of corpora and to create a final product.

As an explanation for why this method works and can be at least a good complement to other approaches based on lexico-syntactic patterns, we refer to the work of Eco (1979), who described how authors add relevant conceptual features when they introduce concepts in their texts without always doing it explicitly – that is, without patterns such as ***X** is a kind of **Y*** – but rather using appositions and a variety of other textual strategies, such as in the following example, where the author is stating a hyperonymy relation between 'mammal' on the one hand and 'platypus' and 'spiny anteater' on the other: *The **platypus** and the **spiny anteater**, the only two **mammals** that lay eggs, were first described in* [...].

Among the lines of future work, we will perform more extensive evaluations with more words, from different semantic groups, and even in different languages. A word-sense disambiguation strategy would undoubtedly benefit precision, and we also need to account for the study of multiword expressions, which were not included in this paper. Furthermore, in the development of a practical application, we should also consider the possibility of taking user feedback into account, in such a way that the accuracy of the algorithm could improve with a step-by-step intervention of a user to correct errors before they reproduce.

Note

[1] This research has been made possible thanks to funding from projects 'Agrupación semántica y relaciones lexicológicas en el diccionario', lead researcher J. DeCesaris (HUM2009-07588 / FILO) and APLE: 'Procesos de actualización del léxico del español a partir de la prensa', lead researcher: M.T. Cabré (FFI2009-12188-C05-01 / FILO). The authors would like to thank the anonymous reviewers for their help and Mark Andrews for proofreading.

References

Alshawi, H. 1989. 'Computational Lexicography for Natural Language Processing.' In B. Boguraev and T. Briscoe (eds.), *Analysing the Dictionary Definitions.* White Plains, NY: Longman Publishing Group, 153–169.

Chodorow, M., R. Byrd and G. Heidorn 1985. 'Extracting Semantic Hierarchies from a Large On-line Dictionary.' In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 299–304.

Church, K. & Hanks, P. 1991. 'Word Association Norms, Mutual Information and Lexicography.' *Computational Linguistics* 16.1: 22–29.

Dolan, W., L. Vanderwende and S. Richardson 1993. 'Automatically Deriving Structured Knowledge Bases from On-line Dictionaries.' In *Proceedings of the First Conference of the Pacific Association for Computational Linguistics (Vancouver, Canada)*, 5–14.

Eco, U. 1979. *Lector in fabula: la cooperazione interpretativa nei testi narrativi.* Milan: Bompiani.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Firth, J. R. 1957. 'A Synopsis of Linguistic Theory 1930-1955.' *Studies in Linguistic Analysis.* Oxford: Philological Society.

Fox, E., J. Nutter, T. Ahlswede, M. Evens and J. Markowitz 1988. 'Building a large thesaurus for information retrieval.' In *Proceedings of the 2nd Conference on Applied Natural Language Processing, Morristown, NJ, USA*, 101–108.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Construction.* Dordrecht, The Netherlands: Kluwer.

Hanks, P. 2004. 'Corpus Pattern Analysis.' In G. Williams and S. Vessier (eds.), *Proceedings of the Eleventh EURALEX International Congress.* Lorient: Université de Bretagne, 87–97.

Harris, Z. 1954[1985]. 'Distributional Structure.' In J. J. Katz (ed.), *The Philosophy of Linguistics.* New York: Oxford University Press, 26–47.

Hearst, M. 1992. 'Automatic Acquisition of Hyponyms from Large Text Corpora.' *Proceedings of the 14th International Conference on Computational Linguistics (Nantes, France)*: 539–545.

Kress, G. (ed.) 1976. *Halliday: System and Function in Language*. Oxford: Oxford University Press.

Miller, G. A. 1995. 'WordNet: A Lexical Database for English'. *Communications of the ACM*, 38.11: 39–41.

Nazar, R. 2010. *A Quantitative Approach to Concept Analysis*. PhD Thesis, IULA, Universitat Pompeu Fabra.

Nazar, R., J. Vivaldi and L. Wanner. Forthcoming. 'Automatic Taxonomy Extraction for Specialized Domains Using Distributional Semantics.' *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*.

Phillips, M. 1985. *Aspects of Text Structure: An Investigation of the Lexical Organization of Text.* Amsterdam: North-Holland.

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

Lin, D. 1998. 'Automatic retrieval and clustering of similar words.' In *Proceedings of COLING-ACL*, 768–774.

Renau, I. and R. Nazar 2011. 'Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis.' In *Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing.* Huelva: University of Huelva.

Rydin, S. 2002. 'Building a hyponymy lexicon with hierarchical structure.' In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition. Association for Computational Linguistics (Morristown, NJ, USA)*, 26–33.

Schmid, H. 1994. 'Probabilistic Part-of-Speech Tagging Using Decision Trees.' In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.*

Schütze, H. and T. Pedersen 1997. 'A Co-occurrence-based Thesaurus and two Applications to Information Retrieval.' *Information Processing and Management* 3.33: 307–318.

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

Snow, R., D. Jurafsky and A. Ng 2006. 'Semantic Taxonomy Induction from Heterogeneous Evidence.' In *Proceedings of the 21st International Conference on Computational Linguistics (Sydney, Australia)*, 801–808.

Vossen, P. (ed.) 1998. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic.