Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 67-74 recibido 20-04-12 revisado 29 -05-12 aceptado 05-06-12

Co-occurrence Graphs Applied to Taxonomy Extraction in Scientific and Technical Corpora*

Grafos de coocurrencia aplicados a la extracción de taxonomías en corpus científico-técnicos

Rogelio Nazar(1) **Jorge Vivaldi**(1) **Leo Wanner**(2)
(1) University Institute for Applied Linguistics
(2) Department of Information and Communication Technologies
and Catalan Institution for Research and Advanced Studies (ICREA)
Universitat Pompeu Fabra, C/ Roc Boronat, 138, 08018 Barcelona
{rogelio.nazar; jorge.vivaldi; leo.wanner}@upf.edu

Resumen: Los grafos de coocurrencia léxica han sido utilizados en lingüística computacional en experimentos de desambiguación de sentidos pero hasta ahora no para la extracción de relaciones de hiperonimia, donde la metodología más usual ha sido la aplicación de patrones léxico-sintácticos. En este artículo mostramos que es posible extraer relaciones de hiperonimia entre términos utilizando estadísticas de coocurrencia. La clave del método reside en que las relaciones de coocurrencia no suelen ser simétricas en el caso de las relaciones de hiperonimia y, en consecuencia, es posible generar grafos dirigidos de coocurrencia que guardan una apariencia similar a la de una taxonomía. En el presente artículo presentamos experimentos con textos de la Wikipedia en castellano ordenados aleatoriamente, pero los resultados sugieren que la coocurrencia asimétrica entre términos es una propiedad intrínseca y macroscópica del discurso argumentativo en general.
Palabras clave: Construcción de ontologías; estadísticas de coocurrencia; extracción de taxonomías; lingüística cuantitativa; semántica distribucional.

Abstract: Word co-occurrence graphs have been used in computational linguistics mainly for word sense disambiguation and induction, but until very recently, not for the extraction of hypernymy relations, where the methodology most often applied is the use of lexico-syntactic patterns. In this paper, we show that it is possible to use word co-occurrence statistics to extract IS-A relations between entities in scientific and technical corpora. We exploit the fact that word co-occurrence often has a direction, that is, a term might co-occur with another, but this is very often not true the other way round. This means that one can represent co-occurrence as a directed graph and this graph resembles a taxonomy. In this paper we present an experiment with texts randomly extracted from the Spanish Wikipedia, but our findings suggest that this co-occurrence behavior is a macroscopic and intrinsic property of argumentative discourse in general.
Keywords: Co-occurrence statistics; distributional semantics; ontology learning; quantitative linguistics; taxonomy extraction.

1 Introduction

Having at our disposal software that could automatically induce structured data such as a taxonomy of concepts from unstructured text would be, without any doubt, a substantial improvement to our ability to conduct scientific research[1]. Presently, the corpus of scientific literature has reached such a volume that it is becoming increasingly difficult for a single individual or a group of researchers to follow related work and spot all relevant advances in their fields. The desire to obtain structured data from texts has motivated decades of efforts in the area of automatic information extraction (Harris, 1958;

* This research was funded by project APLE (Spanish Ministry of Science and Education: Ref. FFI2009-12188-C05-01/FILO) lead by Dr. M. Teresa Cabré. The authors would like to thank the anonymous reviewers and Chris Norrdin for proofreading.

[1] The paper is based on a chapter of the first author's PhD thesis (Nazar, 2010).

Grishman, 1997). In this paper, we focus on the single case of the extraction of hypernymy relations between scientific concepts, and present a novel approach to the topic from the perspective of term co-occurrence statistics.

The vast majority of the proposals produced so far (see next section) have opted for the so called "pattern-based" approach to taxonomy extraction, in which the strategy is language dependent and consists of generating lists of lexico-syntactic patterns that presumably convey a hypernymy relation as in the case, for instance, of the following pattern: **X** *is a kind of* **Y**. The limitation of this approach is that lexical patterns not always convey the expected relation and, moreover, hypernymy relations often appear in the corpus expressed in ways that the researcher was not able to anticipate.

The present proposal is different from previous work from the starting point. It disregards the use of explicit lexico-syntactic patterns in favor of an "immanent" approach, that is, a corpus-based approach with no previous conceptions about the language or the domain. Another characteristic of the present proposal is that, because our starting point is not based on patterns, our approach is entity-by-entity based, which means that given a set of input terms the output will be the assignment of the most probable hypernym for each term.

A basic sketch of our reasoning can be easily grasped by considering the following statements, assuming that A, B, C, D, E and F are terms that denote real or conceptual entities in a given language and domain, and when we say that one term has a tendency to co-occur with another we mean a significant frequency of co-occurrence of these terms in a given context window, such as a search engine's snippet. Thus, if:

- A tends to occur with B and C
- B tends to occur with C and D
- C tends to occur with D
- D tends to occur with E and F

then, we assume that B and A are hyponyms of C, while C, in turn is a hyponym of D. Naturally, conclusions of this kind are not based on a small number of cases as in this sketch, but on hundreds of contexts of occurrence of the input terms found in a corpus. A few heuristics, described in detail in Section 3, are added to the procedure. The first is the operational definition of what counts as an entity. For simplicity, terms denoting entities are not selected observing authoritative sources that certify their normative status. Instead, they are selected according to a statistic criterion as word n-grams which show a significant frequency of occurrence. This procedure is overly simplistic and causes a certain amount of noise in the results. It is to be expected that strategies based on syntactic chunking could offer room for improvement in this aspect, these strategies being linked to the fields of Terminology Extraction (TE) and Named Entity Recognition (NER). In any case, it should be clear that the purpose of the present paper is not to offer the best possible results in the extraction of taxonomy links, a task that would demand a combination of different methods and resources, but to test how much can be done using only limited and essentially simple information such as co-occurrence statistics and, eventually, elementary inferences from these data.

2 Related Work

The fields of TE and NER have been evolving independently of efforts in automatic conceptual relation extraction, but are relevant to all methods of taxonomy extraction because before any attempt to extract relations between terms can be undertaken, these terms must be defined in some way. Lack of space prevents us from offering a detailed introduction which would refer to other works on TE (Kageura & Umino, 1996; Vivaldi & Rodríguez, 2011; Nazar, 2011) and NER (Grishman & Sundheim, 1996; Nadeau & Sekine, 2007).

With respect to efforts in automatic taxonomy extraction, reports began to appear shortly after the availability of the first copies of digitalized lexicographic material, sharing the point of view and methodology: of crafting a specific script for each dictionary and processing the definitions identifying the head of the defining phrase as the hypernym candidate (Amsler, 1981; Chodorow et al, 1985; Fox et al, 1988; Alshawi, 1989). These rules are written in the form of lexico-syntactic patterns and can capture not only hypernymy relations but others such as Part-of, Object-of, Location, Purpose, Manner,

Size, Time, Agent, Act-of, Set-of, Inhabitant-of, Follower-of, etc.

When corpus linguistics gained momentum, at the beginning of the nineteen nineties, researchers in the area started to try to derive taxonomies directly from corpora instead of dictionaries. However, the underlying methodology and assumptions were essentially the same as in previous attempts with machine readable dictionaries (Hearst, 1992; Pearson, 1998; Morin, 1999; Meyer, 2001; Rydin, 2002; Auger & Barrière, 2008). Under the influence of these authors, researchers conducting studies on the subject of conceptual relation extraction will typically define first a corpus of a domain to work with and then apply a routine that searches through the space of this corpus for any occurrence of the members of a set of handcrafted patterns with the aid of a concordance extraction tool. Then, for each context of occurrence, they will inspect what kind of entities are at each side of the pattern and, if these entities really hold the desired conceptual relation, then the outcome will be considered a success.

Few works depart from this perspective. There are reports on the use of machine learning techniques to automatically extract the lexico-syntactic patterns, saving thus the effort of creating them manually (Snow et al, 2006). Patterns are learned with the help of seed-patterns or seed term-pairs which instantiate the relation in question, which is gradually expanded with similar instances found in a corpus. Other authors have proposed the use of statistical techniques to find semantic similarities between entities, inducing vector-based thesauri (Grefenstette, 1994; Lin, 1998). The reasoning is that entities which can be classified as, say, beverages, have a distinctive distributional similarity (e.g., *a bottle of X*, *drinking too much X*, etc.).

3 Methods

As already mentioned in the introduction, our approach to taxonomy induction from corpora is based on statistics of term co-occurrence. The context window –or the space for that co-occurrence– is a paragraph of text (in practice, the text between two newline characters). Using a large corpus, we have been able to observe how hypernymy relations are correlated with term co-occurrence and, given the asymmetric property of term association, taxonomic links can be automatically derived without explicit linguistic or ontological knowledge. A description of each of the steps of the experiment follows. Among them, the most important are the study of first order co- occurrence or syntagmatic association (Section 3.3), the study of second order co-occurrence or paradigmatic relation (Section 3.4) and, finally, the representation of these co-occurrence relations in a directed graph (Section 3.5).

3.1 Selection of a sample of terms to be used as input

The idea of analyzing terms in batches of hundreds instead of one-by-one will become apparent in the following subsections. The basic motivation is that the algorithm needs large numbers to obtain reliable estimations. Given a set of input terms in the same language and domain, each will result in the assignment of the most probable hypernym. In practice, these terms can be obtained from a glossary or database or be the result of an extraction of terms from a corpus.

3.2 Compilation of a reference corpus of the language

A reference corpus is needed in order to develop a language model that will allow us to score and highlight the most significant terms. This model consists of the frequencies of occurrence of words and word n-grams in a corpus of general language. A corpus of press articles of an extension of two million words is sufficient to be used as a language model. Of course, more data would produce better results.

3.3 Analysis of first order co-occurrence

The analysis of first order co-occurrence consists in extracting terms that are syntagmatically related to an input term, which is done by sorting the co-occurring vocabulary in decreasing order of frequency. For illustration, consider an example in the field of medicine (Table 1). These are the most frequent n-grams in the first 100 snippets returned by a web search engine using the term *chronic obstructive pulmonary disease* (COPD).

This co-occurring vocabulary is defined as a set of n-grams ($n \leq 3$) with term frequency and document frequency ≤ 3. Units of a

Rank	Term	Freq.
1	copd	45
2	disease	23
3	lung	21
4	lung disease	20
5	chronic bronchitis	18
6	chronic	18
7	chronic obstructive	14
8	bronchitis and emphysema	12
9	emphysema	10
10	known as copd	9
11	copdgroup of lung	9
12	obstructive	8

Table 1: Terms that most frequently co-occur with *chronic obstructive pulmonary disease*.

length of less than 4 characters and multiword units with a first or last element of a length less than 4 characters are eliminated. The rest of the vocabulary is weighted in order to keep only those units that show a significant frequency. The weight is calculated as shown in (1), where i is the n-gram from the frequency lists, $f_o(i)$ the observed relative frequency of i in the analyzed corpus and $f_e(i)$ the expected relative frequency of i, which is its frequency in the reference corpus. We eliminate all units with a score above an empirically determined threshold (0.01), as well as all n-grams with a first or last element in the same condition.

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

If a lexical resource for the analyzed language is available, it can be used at this point to filter out units that are not nouns and to change plural forms into their lemmata. In case this linguistic resource is not available, a workaround can be to proceed with a pseudo-lemmatization based on orthographic similarity using a similarity coefficient such as Dice (2) with letter bigrams as features.

$$Dice(I,J) = \frac{2|I \cap J|}{|I| + |J|} \quad (2)$$

In order to avoid the possibility of two components of the same n-gram competing for positions in the rank, we eliminate overlapping units. That is, if a unit forms part of another and both have the same frequency, such as in the case of *chronic* and *chronic bronchitis*, it means that every time *chronic* occurs, it is followed by *bronchitis*. In such cases, only the n-gram with higher n is kept. The absolute frequency of the remaining n-grams is multiplied by their corresponding n.

3.4 Analysis of second order co-occurrence

The analysis of second order co-occurrence is very similar to the first order, the only difference being that it is the re-iteration of the analysis for each of the terms that were found co-occurring with the input term. Thus, if on the first analysis for the term *chronic obstructive pulmonary disease* we found that it is related to *disease* and *lung disease*, etc., now we will submit these new terms to the same process. The result is that, for the initial term *chronic obstructive pulmonary disease*, we will find terms that are related to it, and also terms that are related to these latter terms. That enables us to calculate a dispersion coefficient $D(i,j)$ which measures how recurrent a term j is among the lists of related terms generated from a term i. The rationale is that the correct hypernym term j of a term i not only appears syntagmatically related to i, but it is also related to other co-hyponyms that co-occur with i. This dispersion is calculated by multiplying the observed frequency of a hypernym candidate with the number of times the candidate appeared in the frequency lists, as shown in (3), where $f_o(i,j)$ is the observed frequency of j in the contexts of i and all of its related terms, and $D(i,j)$ is the dispersion of term j in the analysis of i. Table 2 shows the terms that have a significant second-order co-occurrence with the initial input term.

$$wD(i,j) = \log(1 + f_o(i,j) * D(i,j)) \quad (3)$$

Rank	Term	1st ord	2nd ord
1	disease	23	142
2	symptoms	7	134
3	chronic	18	77
4	pain	124	3
5	causes	86	3
6	lung	21	63
7	chronic obstructive	7	36
8	lung disease	10	30
9	chronic bronchitis	9	28

Table 2: Terms that show high frequency of second order co-occurrence with *chronic obstructive pulmonary disease*.

3.5 Generation of a directed graph of term co-occurrence

Once the first and second order co-occurrence has been calculated for a sample of terms, the next step is to establish the hypernymy relations between the terms. The central criteria for the selection of a term j as a hypernym of the term i is that j is in first or second order co-occurrence with i and that j is in the same situation with respect to other input terms. This is represented as $wT(j)$ in (4): the number of times j appears in the hypernym candidate lists of other terms in the input term list, with H as the hypernym list of term i, H_i as the hypernym list of $i \in H$, and $|j \in H_i|$ as the number of times j occurs in H_i).

$$wT(j) = \sum_{i=1}^{|H|} |j \in H_i| \qquad (4)$$

With the hypernym j of i determined, it can be assumed that all members of the candidate list of i that also have j as the most generic term are hypernyms of i. Due to the transitivity property, it is also possible that i ends up as hyponym of a term that is not syntagmatically related. In any case, the result of the process is that for each input term the system will attempt to return one term as the best hypernym candidate available.

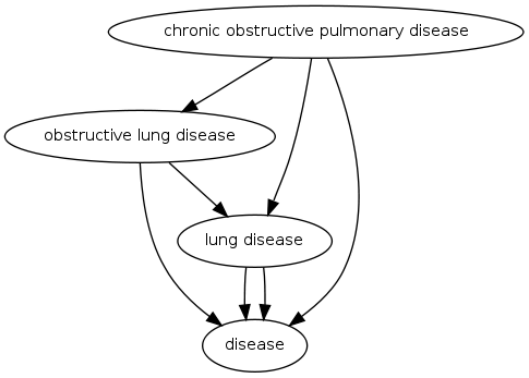

Figure 1: Example of a co-occurrence graph resembling a taxonomy chain for the term *chronic obstructive pulmonary disease*

Figure 1 shows that COPD has a tendency to co-occur with the terms *obstructive lung disease*, *lung disease* and *disease*, but none of these three show a special tendency to co-occur with COPD (the relation is not reciprocal). The term *obstructive lung disease*, in turn, shows a similar tendency to co-occur with *lung disease* and *disease*, but again, none of these two selects *obstructive lung disease* as a frequent co-occurrence. Finally, *lung disease* only co-occurs with *disease* and this last term does not show a tendency to co-occur with any of the above. Coding these relationships as arrows in the graph, we can interpret the figure as a taxonomy, having the number of incoming arrows as a natural expression of hypernymy. The interpretation is that COPD is a kind of disease.

3.6 Analogical inference

Working with corpus based methods has the shortcoming that very often the terms we are analyzing are not found in the corpus with sufficient frequency. In order to overcome this limitation, we have extended the co-occurrence model with the addition of a layer of analogical inference.

In essence, the idea is that if a given term is not found in the corpus, we will attempt to find some similarity with other terms which have been found in the corpus and effectively assigned a hypernym. For instance, when the algorithm has found that in repeated occasions terms such as *lung disease* or *celiac disease* have the term *disease* as hypernym, then it can safely assume that another term that was not found in the corpus but has clear similarities such as, for instance, *Knights disease*, is also another kind of disease. Notice that this cannot be computed by a simple overlapping measure, i.e., we cannot assume that *Knights disease* is a *disease* just because the word *disease* is included in the term *Knights disease*. Doing so would also lead us to other wrong assumptions, for instance that *lichen planus* is a kind of *planus* or that a *transforming growth factor* is a kind of *factor* when actually they are a kind of *disease* and a kind of *protein*, respectively. The procedure is, thus, not just to find overlapping sequences but to learn to associate features in the terms with the hypernyms they have been assigned by the co-occurrence method. This reasoning allows us to operate in the same way in cases where there is actually no overlapping. For instance, when the algorithm finds that terms such as *Carpenter syndrome* and *Asperger syndrome* consistently receive the term *disease* as hypernym, it will assume that other terms that were not found in the corpus but share the same element *syndrome*, such as *Meretoja syndrome* or *Maffuci syndrome*,

can also be considered diseases by means of simple analogy.

The same reasoning is applied at the morphological level, which in this case is defined as the first and last four letters of each word of a term. The motivation behind this procedure is that very often in specialized terminology the units that pertain to a particular semantic class share some morphological features. Thus, if the algorithm finds a persistent morphological pattern within the terms of a determined semantic class, it will learn to associate such pattern with that class. For example, if it finds that terms ending with a sequence of letters such as *-itis* or *-osis* which are frequently assigned the hypernym *disease*, as in the case of *arthritis* or *endometriosis*, then it will assume that other terms that were not found in the corpus such as *acrodermatitis* or *pneumocystosis* can also be classified as diseases.

The main benefit of this analogical inference is that it grants the algorithm a great amount of flexibility and generalization power, because there is no need for explicit information about the entities nor any kind of previous training phase. The learning is conducted on the fly using the result of the co-occurrence method.

4 Results and Evaluation

In order to evaluate the strategy, we took a sample of 375 terms in Spanish from the Mosby (2003) dictionary, pertaining to the classes of bones, disorders, ganglia, glands, hormones, drugs, organs, proteins and viruses in unequal proportions. This facilitates the task of evaluation because we know that the correct hypernym of each input term must pertain to some of these classes (which is information that the algorithm does not have). As mentioned earlier, in a real life scenario these input terms would be obtained, for instance, by term extraction from LSP corpora. In this experiment we used texts from the Spanish Wikipedia of the year 2010 as corpus, in random order and excluding, of course, all metadata and structural information that could be used to construct a taxonomy, leaving a single text file of approximately 455 million tokens. The choice of this corpus does not mean that we are particularly interested in Wikipedia. In fact, we believe the experiment could be replicated with any other corpus if it is large and diverse enough to contain the input terms with sufficient frequency. It is probable that better results could have been obtained using the web as corpus, but then there would be other factors that we would not be able to quantify, such as the particular ranking of results provided by each search engine.

Only 164 of the 375 input terms appeared with sufficient frequency to select co-occurring terms. Those which did appear in the corpus were assigned a hypernym-candidate, which in the majority of the cases was indeed a correct hypernym and when it was not, it was a semantically related concept. Table 3 shows the evaluation figures for all the experiments. Results are reported both with the co-occurrence method as explained in Sections 3.1. to 3.5. and including the analogical inference layer described in Section 3.6. As we can see, the inference layer dramatically increases figures of recall. This is of course important for a practical application, but the key point that we wanted to demonstrate with this experiment is that asymmetric co-occurrence by itself is sufficient to show a significant correlation with hypernymy relations.

It should also be noticed that the task of assigning a hypernym to a given term is performed with variable levels of precision depending on the domains. In the case of the terms pertaining to the class of ganglia, the system found virtually no occurrence of them in the corpus, and this explains why there are so many zeros in their case (they are false negatives). However, because of their extremely regular form – most of them happen to be terms such as *ganglio intercostal* (intercostal node), *ganglio inguinal superficial* (superficial inguinal node), *ganglio gástrico* (gastric ganglion), and so on – this makes it possible for the inference engine to assign a correct hypernym *ganglio* (node/ganglion) to all of them. There are also many zeros in the case of organs, but this time for a different reason. They did appear but the performance was extraordinarily poor because organs are not associated with the hypernym *órgano* (organ) but, instead, to a meronym like *cuerpo* (body) and are, therefore, false positives. Something similar occurs with organs that are related to diseases, e.g. *próstata* (prostate) as a type of cancer.

		Co-occurrence						Co-occurrence + Inference					
Domain	**Trials**	**tp**	**fp**	**fn**	**P**	**R**	**F1**	**tp**	**fp**	**fn**	**P**	**R**	**F1**
Bones	36	27	1	8	96.43	75	84.38	32	1	3	96.97	88.89	92.75
Disorders	68	27	3	38	90	39.71	55.10	56	2	10	96.55	82.35	88.89
Ganglia	29	0	0	29	0	0	0	29	0	0	100	100	100
Glands	15	3	3	9	50	20	28.57	15	0	0	100	100	100
Hormones	43	25	7	11	78.13	58.14	66.67	31	7	5	81.58	72.09	76.54
Drugs	69	4	4	61	50	5.80	10.39	27	7	35	79.41	39.13	52.43
Organs	29	0	22	7	0	0	0	0	22	7	0	0	0
Proteins	65	15	15	35	50	23.08	31.58	31	24	10	56.36	47.69	51.67
Virus	21	2	6	13	25	9.52	13.79	15	6	0	71.43	71.43	71.43
TOTAL	375	103	61	211	62.80	27.47	38.22	236	69	70	77.38	62.93	69.41

Table 3: Results after 375 experiments.

In this case, there is nothing the inference engine can do, because it does not rectify the output of the co-occurrence method. Other causes of mistakes have been the wrong segmentation of multi-word terms and also errors due to problems of polysemy.

We cannot offer a comparison to other authors' results because it is technically impossible given the fact that our method is entity based, thus there is no way of replicating a pattern based method using our dataset. In any case, compared with state-of-the-art techniques, an F1 of %69 is not a very impressive result. In fact, we can expect to obtain higher precision figures with a very simple baseline, such as taking the lexical unit that is the head of the phrase in multiword terminology (normally the first noun from the left in the case of Spanish noun phrases). That said, our results are more meaningful than those that can be obtained with a trivial baseline: undoubtedly, *bocio* (goitre) is the correct hypernym of *bocio coloide* (colloid goiter), and *citocromo* (cytochrome) is of *citocromo P-450* (cytochrome P450), but for most NLP applications, hypernyms such as *enfermedad* (disease) and *enzima* (enzyme) are more meaningful. This is one of those cases in computational linguistics where the interesting point is not to have obtained figures representing better precision than other methods but to have found a methodology to extract information that could not be obtained otherwise.

5 Conclusions and Future Work

This paper has presented an experiment in taxonomy extraction from corpus using purely statistical methods. Our approach to the topic is fundamentally theoretical, though based on empirical evidence. We believe we have found a measurable pattern of term co-occurrence which is characteristic of hypernymy relations, and this property is expected to be inherent to argumentative discourse independent of the language and the domain. Still, much experimentation has to be carried out before reaching conclusive results, especially to sustain the claim of language independence. However, and despite our theoretical motivation, nothing seems to prevent an application of this algorithm as a method for automatic development of taxonomies from corpora at least in the languages where it has already been tested (English and Spanish, so far).

As an interpretation of why this method yields results, we recall two discursive strategies that can be identified in scientific or argumentative discourse. One of the strategies is to introduce and define concepts in discourse according to the knowledge established in the community targeted by the text. To present something new is the purpose of the other strategy used in the text. In this case, the statements convey external or empirical information which cannot be directly inferred from the established knowledge (or at least not trivially). One of the consequences of these two forces acting upon discourse is that we can expect a certain degree of coincidence in the passages where authors introduce concepts in their texts, and this coincidence can be measured in the selection of relevant conceptual features, often hypernym terms.

Future work will include replicating the same model in different languages and domains and introducing different degrees of explicit linguistic knowledge such as POS-tagging, chunking, lexico-syntactic patterns and also ontologies and other semantic re-

sources. We expect to automatically produce high quality taxonomies in the near future with a combination of different techniques.

References

Alshawi, H. 1989. Computational lexicography for natural language processing. *Analysing the dictionary definitions.* Longman Publishing Group, (White Plains, NY, USA): 153–169.

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

Auger, A. & Barrière, C. 2008. Pattern-based Approaches to Semantic Relation Extraction Special issue of Terminology. *Terminology* 14(1).

Chodorow, M. & Byrd, R. & Heidorn, G. 1985. Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd annual meeting on Association for Computational Linguistics, July 08-12, 1985 (Chicago, Illinois, USA): 299–304.

Fox, E. & Nutter, J. & Ahlswede, T. & Evens, M. & Markowitz, J. 1988. Building a large thesaurus for information retrieval. Proceedings of the 2nd conference on Applied natural language processing, Morristown, NJ, USA. Association for Computational Linguistics: 101–108.

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

Grishman, R. 1997. Information Extraction: Techniques and Challenges. *Information Extraction*, ed. Maria Teresa Pazienza, Springer-Verlag.

Grishman, R. & Sundheim, B. 1996. Message Understanding Conference 6: a brief history. Proceedings of the 16th International Conference on Computational Linguistics (Copenhagen, Denmark): 466-471.

Harris, Z. 1958. Linguistic transformations for information retrieval. Proceedings of the 16th International Conference on Scientific Information. National Academy of Sciences-National Research Council (Whashington DC, USA).

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

Kageura, K. & Umino, B. 1996. Methods of Automatic Term Recognition. *Terminology*, 3(2): 259–290.

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

Meyer, I. 2001. Extracting knowledge-rich contexts for terminography. Recent Advances in Computational Terminology.

Morin, E. 1999. Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. PhD Thesis in Computer Sciences, Université de Nantes, 1999.

Mosby 2003. Diccionario Mosby de medicina, enfermería y ciencias de la salud. VI Edición. Madrid: Elsevier.

Nadeau, D., Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26.

Nazar, R. 2010. A Quantitative Approach to Concept Analysis. PhD thesis, Universitat Pompeu Fabra.

Nazar, R. 2011. A Statistical Approach to Term Extraction. *International Journal of English Studies* 11(2):153-176.

Pearson, J. 1998. *Terms in context.* John Benjamins.

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

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

Vivaldi, J.; Rodríguez, H. 2011. *Extracting Terminology from Wikipedia. Procesamiento del lenguaje natural* 47: 65–73.