A Method for Extracting Hyponymy-Hypernymy Relations from Specialized Corpora using Genus Terms

Olga Acosta[α], César Aguilar[β] and Gerardo Sierra[γ]

[α]Postgraduate School of Computer Science, UNAM, Ciudad Universitaria, Mexico City, Mexico
[β]Department of Linguistics, Autonomous University of Queretaro, Queretaro, Mexico
[γ]Language Engineering Group, Engineering Institute, UNAM, Ciudad Universitaria, Mexico City, Mexico
{OAcostaL, CAguilar, GSierraM}@iingen.unam.mx
http://www.iling.unam.mx

Abstract. This work exposes a method for extracting hyponymy-hypernymy relations from definitions situated on specialized texts in Spanish. The method starts with the automatic extraction of a set of hyponyms and hypernyms from analytical definitions. The set of hypernyms is employed as a *seed* to extract an additional set of relations from a domain-specific corpus. Our method considers a phase of shallow parsing for extracting fragments of analytical definitions where term and Genus Term are located. This method reaches an average recall over 89% and an average precision of 92% about relations found in a specialized corpus.

Keywords: lexical relation, hyponymy-hypernymy, definitional contexts, shallow parsing, bootstrapping, term and Genus Term.

1 Introduction

The possibility of extracting lexical relations in text corpora is one of the current interests in NLP areas such as computational lexicography and terminology. These areas have been focused mainly on the identification of the hyponymy-hypernymy relation, considering this relation is represented by the canonical sequence Genus Term + Differentia, according to the analytical definition theory formulated by Aristotle.

Based on the use of machine readable dictionaries (MRDs), [1] and [2] made experiments for finding hyponyms and hypernyms inserted in definitions. An important advance in this kind of experiments is the work of [3], where these authors proposed the *IS-A* operator as a way to extract and categorize lexical items (that is, terms and Genus Terms) linked in a relation of hyponymy-hypernymy.

Following these works, the experiment made by [4] offers a significant method for identifying lexical-syntactic patterns associated to hyponyms and hypernyms in large-corpora. However, although patterns of [4] have high precision, [5] and [6] have observed they are rarely used, even when a big corpus is analyzed. So, there are other alternative approaches based on the method developed by [4]:

- **Clustering:** in this approach distribution of context in a corpus is considered. This approach was proposed by [7] and [8].
- **Finding patterns using the WEB**: in this approach are employed new characteristic patterns of the hyponymy-hypernymy relation, taking into account the use of the Web as a huge source of textual information. This approach was proposed by [9]. In Spanish, [10] made an experiment to find new patterns of the hyponymy-hypernymy relation in the WEB.
- **Machine learning:** Finally, an approach for finding hyponymy-hypernymy relations was developed by [11], considering the application of machine learning methods, oriented to recognize useful patterns employing dependency paths.

In line with all these authors, the main goal of our work is the extraction of hyponymy-hypernymy relations from Spanish texts, particularly specialized documents, according to the argumentation formulated by [12] and [13], in order to obtain specialized conceptual information encoded through lexical relations.

To reach this goal, we propose a method that takes advantage of conceptual information extracted from analytical definitions, considering the association established between the term defined and its Genus Term. We obtained these analytical definitions from definitional contexts (DCs) in Spanish, based on the methodology developed by [14] and [15].

Once identified these DCs, we extracted term and Genus Term. We use the hyperonym subset in a bootstrapping step for finding new hyponyms in a specialized corpus. In this phase, prepositional phrase with the preposition *de* (Eng. of/from) is used as a link for finding new instances of hyponyms.

We organize our paper as follows: in the Section 2 we define the notion of conceptual information, focusing its insertion in specialized documents. In addition, we briefly describe how it is possible to extract analytical definitions from DCs obtained from specialized texts and analyze the role of the preposition *de* in the hyponymy-hypernymy relation between terms and Genus Terms. In the Section 3 we describe our methodology for automatic extraction of lexical relations. Then, in the Section 4 we expose our evaluation and the results that we obtained after to apply our method. Finally, in the Section 5 we formulate our conclusions and propose future works.

2 Conceptual information

Nowadays, computational lexicography and terminology are able to recognize concepts in large-text corpora. For solving this recognition, it is important to establish what the best source for obtaining relevant concepts is. In this sense, [16] and [17] point out the value of scientific and technical literature as a source to obtain such concepts. In particular, [16] considers definitions as linguistic representation of concepts, because definitions synthesize all the conceptual information linked to terms circumscribed to a domain-specific knowledge. This conception is close to the observation of [18], because the definitions (in our case, specialized definitions) transcend the particular level of our everyday experiences.

On the other hand, [13] argue about difficulties to find conceptual information in general linguistic corpora or Internet. A practical solution is to explore domain specific corpora, because they contain lexical and conceptual information pertaining to a specific subject matter.

In line with these authors, we consider as conceptual information that information expressed by specialized definitions, particularly in analytical definitions constituted by Genus Terms and Differentia, according to [17] and [3]. These authors have used this kind of definition for searching hyponymy-hypernymy relations established between terms and Genus Terms.

In order to recognize these relations, [3] used the IS-A operator for finding lexical-syntactic patterns with a high degree of precision (e.g.: *an autobiography IS-A kind of book*), particularly in corpora generated from MRDs. However, [14] observes that this kind of patterns is not sufficient for describing all the possibility to express an analytical definition in natural language. Thus, it is necessary to consider other alternative patterns capable to introduce these definitions in specialized documents.

2.1 Definitional contexts extraction

[14] developed a method for extracting terms and definitions in Spanish, which are expressed in textual fragments inserted in specialized documents. These fragments are called definitional contexts (or DCs) and are constituted by a term, a definition, and linguistic or metalinguistic forms, such as verbal phrases, typographical markers and/or pragmatic patterns, for example:

> ***La energía primaria****, en términos generales, se define como aquel recurso energético que no ha sufrido transformación alguna, con excepción de su extracción.* (Eng. The **primary energy**, in general terms, is defined as a resource that has not been affected for any transformation, with the exception of its extraction.)

We can see here a DC sequence formed by the term *energía primaria* (Eng. Primary energy), the definition *aquel recurso…* (Eng. a resource that…) and the verbal pattern *se define como* (Engl. is defined as), as well as other characteristic units such as the pragmatic pattern *en términos generales* (Eng. in general terms) and the typographical marker (bold font) that in this case emphasizes the presence of the term.

For achieving this extraction, [14] employ verbal patterns that operate as connectors between terms and definitions. Such patterns syntactically are predicative phrases (or PrP), configured around a verb that operates as a head of this PrP. Among verbs that work as heads of PrPs, the verb *ser* (Eng. to be) is the most

frequent, mainly because it allows to structure operators such as IS-A. Nevertheless, other verbs can be heads of these PrPs, e. g.: *definir* (Eng. to define), *denominar* (Eng. to denominate), *conocer* (Eng. to know), and others. The following examples show analytical definitions using these verbs:

1. *La [conjuntivitis]* ^{Term} *es una [inflamación]* $^{Genus\ Term}$ *de la conjuntiva del ojo*. (Eng. [Conjunctivitis] ^{Term} is an [inflammation] $^{Genus\ Term}$ of the conjunctiva of the eye).
2. *Se define [conjuntivitis]* ^{Term} *como una [inflamación]* $^{Genus\ Term}$ *de la conjuntiva del ojo*. (Eng. It is defined [conjunctivitis] ^{Term} as an[inflammation] $^{Genus\ Term}$ of the conjunctiva of the eye).

In (1) and (2), we observe terms and analytical definitions linked through PrPs whose heads are the verbs *es* and *define*. In both cases, the term *conjunctivitis* is conceived as an inflammation of the eye, for this reason the Genus Term of these definitions is the term *inflammation*. According to [3], these cases are a canonical example of hyponymy-hypernymy relations into analytical definitions.

We can understand the relation between term and Genus Term as the association of an entity (or the term) which is a member of a set represented by the Genus Term (paradigmatic relation). In analytical definitions, particularly those formulated in technical and scientific texts, we observe that the Genus Term refers to any specific category whose function is to describe and circumscribe the proper attributes of an entity. This categorization can reveal different perceptions of the same entity when a particular category is assigned to it. In the following examples, all these categories capture different perspectives related to the term *diabetes*:

3. *La [Diabetes]* ^{Term} *es una [enfermedad]* $^{Genus\ Term}$ *poligénica caracterizada por niveles anormalmente altos de glucosa en la sangre.* (Eng. [Diabetes] ^{Term} is a polygenic [disease] $^{Genus\ Term}$ characterized by abnormally high glucose levels in the blood).
4. *[Diabetes]* ^{Term} *es la [incapacidad]* $^{Genus\ Term}$ *del cuerpo para producir, o la incapacidad de metabolizar.* (Eng. Diabetes] ^{Term} is the [inability] $^{Genus\ Term}$ of the body to produce, or the inability to metabolize, the human hormone insulin).
5. *[Diabetes]* ^{Term} *es un [desorden]* $^{Genus\ Term}$ *metabólico causado por la producción inadecuada.* (Eng. [Diabetes] ^{Term} is a metabolic [disorder] $^{Genus\ Term}$ caused by inadequate production or utilization of insulin).
6. *[Diabetes]* ^{Term} *es una [condición]* $^{Genus\ Term}$ *en la que una persona tiene un nivel alto de glucosa.* (Eng. [Diabetes mellitus] ^{Term} is a [condition] $^{Genus\ Term}$ in which a person has a high blood sugar (glucose) level).

Additionally, [15] develops an automatic system named ECODE (*Extractor Automática de Contextos Definitorios*), which recognizes, extracts and collects Spanish DCs from specialized texts. For our experiments, we take the ECODE system as a source for obtaining valid candidates of DCs.[1]

2.2 Shallow parsing exploration

Taking into account the regularity of these PrPs as connectors between terms and analytical definitions, we performed an experiment to extract terms and Genus Terms from corpus through of a shallow parsing phase, according to the heuristics proposed by [3]. This shallow parsing phase was programmed employing the NLTK module of Python [19]. Our grammar obtained a recall of 81% of the DCs, and a precision of 98% for recognizing and extracting term and Genus Terms.

2.3 A vision of the Genus Term in analytical definitions framed in the set theory

Genus Term represents a definable criterion for membership in a set, that is, all element having common characteristics to the Genus Term will be a member of the set, e.g., Genus Term *disease* will include everything what can be considered as a disease.

[1]For more details about the ECODE System, it is possible to access through the following WEB Site: http://brangaene.upf.edu/ecode/.

Going into more details, an analytical definition has a Differentia that individualizes any element or subset inserted in a common set respect to others. Thus, the Differentia can be used to delimit smaller sets than those defined by the Genus Term. For example, Genus Term *disease* can be linked with *eye*: *disease of the eye* and several diseases occurring in the eyes can be found: {*Thygesons superficial punctuate keratopathy*, *retinopathy of prematurity*, *glaucoma,* and so on}. If this subset has at least one element, these elements can be considered co-hyponymies of the term defined.

Therefore, we can understand *disease of the eyes* as a kind of disease. We consider prepositional phrases with preposition *de* to link Genus Terms with other elements from a domain-specific corpus, in order to obtain more hyponyms.

2.4 Preposition *of* and Lexical Relations

In tasks of term extraction for Catalan ([20]) and Spanish ([21]), patterns with preposition *de* have been considered as one of the most common for structuring terms in specialized domains. Modifiers of a noun, such as prepositional phrases with *de* or adjectives, often configure more specific terms, e.g.: *syndrome – Down syndrome* (syndrome of Down). [9] proposed a method based on measures of the information theory, as well as linguistic information for building taxonomies of terms from a specialized domain. Therefore, *syndrome* would be considered as hypernymy of *Down syndrome* (hyponymy).

On the other hand, results obtained by [22] in the extraction of meronymy-holonymy relations revealed that patterns such as genitive structure and preposition of + [ART] are characteristic in this relation. Similarly, the results generated by [23] show clusters with structures such as genitive, preposition *of* and verb *have,* represent 53% of the analyzed sentences in the search of meronymy-holonymy relations.
Finally, [24] proposed the use preposition *of* for extracting attributes of concepts from the Web, taking into account a test of *attributehood*, that is:

"the * of the C [is|was]"

Preposition *of*, similarly to its grammatical co-members, tends to be polysemic. However, its high productivity in lexical relations of our interest makes it an important element for tasks such as the automatic extraction of lexical relations.

3 Extraction of hyponymy-hypernymy relations

3.1 Corpus

Our corpus is constituted by a set of documents of the medicine domain, basically human body diseases. These documents were collected from MedLinePlus in Spanish. The size of the corpus is 2.7 million of words.[2]

For our experiment, a set of 841 analytical definitions about diseases were collected from Wikipedia. These analytical definitions were manually recognized considering verbal heads mentioned in subsection 2.1. However, in a future, these analytical definitions will be automatically extracted by the ECODE System.

3.2 Description of the methodology

The methodology proposed is based on exploiting a set of Genus Terms (hypernyms) as a *seedset* in a bootstrapping step. This set is intended to collect an additional number of hyponymy relation instances from a domain-specific corpus. Steps required by the methodology are described below:

- A set of analytical definitions and a domain corpus with tagged POS are input to the algorithm. TreeTagger for Spanish [25] was used in this experiment.
- A shallow parsing phase for extracting relevant fragments containing term and Genus Term. We apply heuristics to obtain fragments of text where elements mentioned above can

[2] For more details about the MedLinePlus in Spanish, it is possible to access through the following WEB Site: http://www.nlm.nih.gov/medlineplus/spanish/

be found. Heuristics discussed by [3] were applied for identifying Genus Terms from relevant fragments. The final result of this step is a set C of Genus Terms.

- A set F_1 of noun phrases was obtained from domain corpus. The regular expression applied in a shallow parsing step for extracting noun phrases is:

<NC|NP>+<ADJ>*<PDEL|PDE><NC|NP>+<ADJ>*

- A set V of potential terms was obtained from set F_1 of noun phrases. We assume that noun phrases divided by preposition *de* are potential terms. For example, the noun phrase *folículo piloso de la pestaña* (Eng. hair follicle of the eyelash) produced two terms: *folículo piloso* and *pestaña*. Frequencies of joint and individual occurrence were determined from the set of noun phrases F_1.
- The set C of Genus Terms was used in a bootstrapping step. This step linked each element of C with each element of V. If the concatenation of elements $c_i \in C$, preposition *de* and $v_j \in V$ is part of the set F_1:

$$C_i + \text{<PDEL|PDE>} + v_j \in F_1$$

Then such elements can be related and considered in a new set R.

- We calculated a normalized Pointwise Mutual Information for each pair of elements in R obtained from above step. This normalized measure PMI was proposed by [26].
- Values of joint frequency and PMI were tested for reaching proper levels of recall and precision.

4 Results

The corpus produced 140,917 noun phrases with prepositional phrases as modifier. Specifically, noun phrases with preposition *de* or *de* + ART represented 79% of total. This situation suggests preposition *de* as an important producer of lexical relations.

4.1 Evaluation

In a first evaluation, instances of the hyponymy-hypernymy relation considering patterns proposed by [4] and [10] were obtained from the domain corpus. The number of relations extracted was 1488. Table 1 shows most frequent hypernyms and the number of instances (germination frequency) extracted by patterns of [4] compared with patterns of [10]. [4] does not report precision and recall measures. However, she mentions quality of the relations found seems high overall, although there are difficulties. This is in line with other researchers considering these patterns have high precision but low recall.

With regard to [10], they ranked the list of hyponyms by applying an iterative evaluation process. They calculate the confidence of instances and patterns in accordance with their association. Precision reported in this work for first 200 instances is below 80%.

Table 1. Number of instances of the Hyponymy-Hypernymy relation

Hypernymy	Hearst'patterns	Ortega'patterns
Symptom	41	31
Disease	22	70
Problem	16	33
Organ	16	17
Disorder	2	15

In contrast, our method using a Genus Term set of 64 elements extracted 4818 instances. Table 2 shows most frequent hypernyms, the number of relations (germination frequency) extracted and percentage of true hyponyms extracted for the first 10 Genus Terms. Additionally, table 2 shows the percentage of true hyponyms is over 80%, which justifies the importance of taking into account Genus Terms as a starting point for extracting new relations. In the same way, Figure 1 compares the number of instances extracted by each approach where it is clear our method extracts at least 3 times more instances than patterns of [4] and [10] together.

Table 2. Number of instances of the Hyponymy-Hypernymy relation

Hypernymy	Germination frequency	% true relations
Symptom	455	96
Absence	277	94
Syndrome	281	90
Disease	246	83
Loss	273	85
Increase	233	94
Signal	215	100
Disorder	234	93
Inflammation	180	95
Infection	150	89

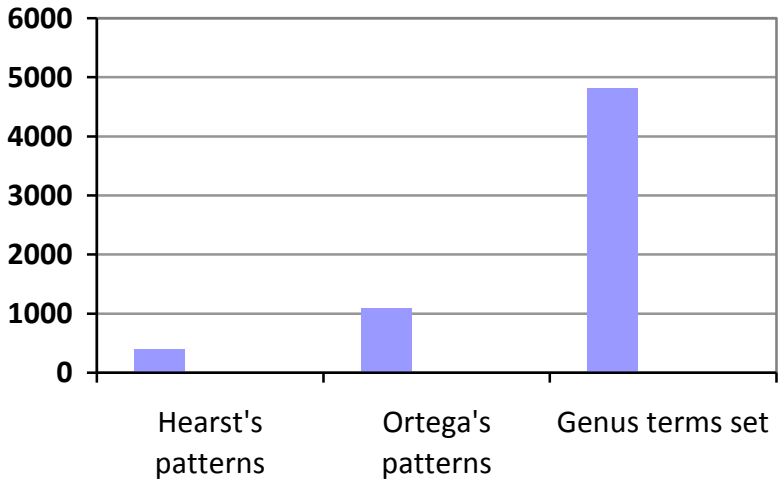

Fig. 1. Comparing instances extracted by the three approaches

Table 3 presents average measures of precision and recall considering two thresholds related with PMI and co-occurrence frequency values. With a PMI>=0.10 we obtained recall levels over 80%. However, according to [26], the mutual information is particularly sensitive to estimates that are inaccurate due to data sparseness. Thus, we tested to use a cutoff and to only look at words with a frequency at least 3. With a PMI>=0.10 and frequency >2 recall was drastically reduced at 18%.

From these results, we assumed Genus Terms to be elements with a high confidence and we removed threshold related with frequency. This decision significantly improved recall without affecting precision. Currently, we are testing to include germination frequency of Genus Terms as an important element for determining better levels of confidence of instances. In the same way, we are collecting more documents in order to increase our corpus and to cope data sparseness problem. In the Table 3 we show the Precision, Recall and F measures for two schemas of thresholds.

Table 3. Measures F, precision and recall

Threshold	F	Precision	Recall
PMI>=0.10 and frequency>2	29%	93%	18%
PMI>=0.10	89%	92%	87%

5 Conclusions

We have proposed an alternative approach to extract hyponymy-hypernymy relations using analytical definitions obtained from DCs. Results indicate our method is able to extract a larger number of instances of hyponyms and hypernyms, in comparison with [4] and [10]. Furthermore, precision and recall obtained with our method is significant compared with approaches mentioned.

We assume conceptual information contained in analytical definitions is useful to improve identification of instances mentioned. In fact, analytical definitions offer clear boundaries related to the conceptual information underlying in specialized texts. Thus, we think that a previous step in a task of lexical relation extraction is the identification of DCs which contain definitions well-formed.

Finally, we are interested in exploring other lexical relations such as meronymy-holonymy and attribution structured around preposition *de*. Currently, heuristics based on derivational morphology and partitives, as well as the structure imposed by preposition *de* are being considered.

Acknowledgments: This paper was made possible by the financial support of the Consejo Nacional de Ciencia y Tecnologia, CONACYT, the DGAPA-UNAM, the School of Computer Science, UNAM, and the Autonomous University of Queretaro. Also, we wish to thank the anonymous reviewers for their comments and suggestions.

References

1. Amsler, R.: A taxonomy for English nouns and verbs. In: Proceedings 19th Annual Meeting of the Association for Computational Linguistics, pp. 133-38. Stanford University, California (1981).
2. Alsawhi, H.: Processing dictionary definitions with phrasal pattern hierarchies. Computational Linguistics 13(3-4), 195-202 (1987).
3. Wilks, Y., Slator, B., Guthrie, L.: Electric Words: dictionaries, computers and meanings. MIT Press, Cambridge, MA (1995).
4. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of COLING-92, pp. 539-545, Nantes, France (1992).
5. Cimiano, Ph., Pivk, A., Schmidt, L., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In: Proceedings of the ECAI 2004 Ontology Learning and Population, Valencia, Spain (2004).
6. Ryu, K., Choy, P.: An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning. In: Buitelaar, P., Cimiano, P., & Magnini, B. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications, 15-28. IOS Press, Amsterdam (2005).
7. Pereira, F., Lee, L., Tishby, N.: Distributional Clustering of English Words. Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 183-190. Ohio State University, Columbus, Ohio (1993).
8. Faure, D., Nedellec, C.: A Corpus-based Conceptual Clustering Method for Verb Frames and Ontologies. Proceedings of the LREC Workshop on adapting lexical and corpus resources to sublanguages and applications, pp. 5-12. Granada, Spain (1998).
9. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. Proceedings of Conference on Computational Linguistics/Association for Computational Linguistics, pp. 113-120, Sydney, Australia (2006).
10. Montes, M., Ortega, R., Villaseñor, L.: Using Lexical Patterns for Extracting Hyponyms from the Web. In: MICAI 2007. Advances in Artificial Intelligence. LNCS, Vol. 4827, pp. 904-911. Springer, Berlin (2007).
11. Snow, R., Jurafsky, D., Ng, A.: Semantic Taxonomy Induction from Heterogenous Evidence. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 801–808, Sydney, Australia (2006).
12. Riloff, E., Shepherd, J. A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction. Journal of Natural Language Engineering, 5(2), 147-156 (2004).
13. Buitelaar, P., Cimiano, Ph., Magnini, B.: Ontology learning from text. IOS Press, Amsterdam (2005).
14. Sierra, G., Alarcon, R., Aguilar, C., Bach, C.: Definitional verbal patterns for semantic relation extraction. Terminology, 14(1), 74-98 (2008).
15. Alarcón, R.: Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios. Ph. D. Dissertation. IULA-UPF, Barcelona (2009).

16. Sager, J. C.: A Practical Course in Terminology Processing. John Benjamins, Philadelphia/Amsterdam (1990).
17. Smith, B.: Ontology. In: Floridi, L. (ed.), Blackwell Guide to the Philosophy of Computing and Information, pp. 155-166. Blackwell, Oxford (2003).
18. Pinker, S.: Words and rules. Weindefeld & Nicholson, London (1999).
19. Bird, S., Klein, E., Loper, E.: Natural Language Processing whit Python. O'Reilly, Sebastropol, CA. (2009).
20. Estopà, R.: Extracció de terminologia: elements per a la construcció d'un SEACUSE. Ph. D. Dissertation, IULA-UPF, Barcelona (2003).
21. Vivaldi, J.: Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas. Ph. D. Dissertation, IULA-UPF, Barcelona (2001).
22. Berland, M., Charniak, E.: Finding parts in very large corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 57-64, Orlando, Florida (1999).
23. Girju, R., Badulescu, A., Moldovan, D.: Automatic Discovery of Part–Whole Relations. Computational Linguistics, 32(1), pp. 83-135 (2006).
24. Poesio, M., Almuhareb, A.: Feature-Based vs. property-based KR: An empirical perspective. In: Proceedings of International Conference on Formal Ontology in Information Systems, Torino, Italy (2004).
25. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, September (1994), www.ims.uni-stuttgart.de/~schmid/.
26. Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Chiarcos, C., Castilho, E., Stede, M. (eds). From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, pp. 31-40, Gunter Narr Verlag, Tübingen (2009).
27. Manning, C., Schütze, H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999).