Extracting Concrete Entities through Spatial Relations

Olga Acosta and César Aguilar

Facultad de Letras, Pontificia Universidad Católica de Chile

{oacostal,caguilara}@uc.cl

Abstract. This paper focuses on the automated extraction of concrete entities from a specialized-domain corpus. Then, in a bootstrapping phase, the candidates are used to extract new candidates. Concrete entities are automatically identified by a set of spatial features. In a spatial scene something is located by virtue of the spatial properties associated with a reference object. The axial properties are represented by place adverbs. Additionally, for identifying referent objects in a sentence we consider syntactical patterns extracted by chunking. In order to reduce noise in results, we take into account a corpus comparison approach and linguist heuristics. Results show high precision in candidates with high weights.

Keywords: Concrete entities, lexical relation, information extraction, term extraction, axial properties, nominalization.

1. Introduction

In recent years, the automatic mining of relevant knowledge in the biomedical domain has become in an interesting research area, particularly in tasks related to the generation of taxonomies and ontologies (Smith and Kumar, 2004). This kind of tasks require the design and implementation of efficient information extraction (IE) methods, capable of identifying and extracting textual patterns that contain such relevant knowledge.

Therefore, in this work we propose a methodology for the automatic extraction of concrete entities implicit in medical documents. Then, in a bootstrapping phase, these candidates are used for extracting a larger set of new candidates.

Linguistically speaking, a main concern is those noun phrases (NP) whose modifiers are relational adjectives and where the noun head is a concrete entity, because relational adjectives introduce semantic features which describe specific properties such as formal, constitutive, telic and agentive qualities (Fábregas, 2007). The identification of this type of NP contributes to delimit the number of possible semantic relations. For testing our method, we work with a corpus of medical texts in Spanish.

We organize our paper as follows: in section 2 we define what a concrete entity is, taking into account the description proposed by Fellbaum (1998) for classifying names in WordNet. Then, in section 3, we show a brief explanation about the representation of space in natural language, according to a cognitive framework. In section 4, we describe the most common deverbal nominalizations in specialized texts. In section 5 we explain the relation noun + relational adjective in order to delineate a set of linguistic heuristics useful for filtering non-relevant adjectives. In section 6 we describe our methodology. In section 7 we offer a description of preliminary results. Finally, in section 8, we give our conclusions.

2. Concrete entities

We understand all that exists in the world as a concrete entity which something can be predicated (in Aristotle's categories: substance). For example, concrete entities can be *artifactual* categories like *vehicles*, *clothing* and *weapons*, or *natural* kinds like *birds*, *fruits* and *vegetables* (Landau and Jackendoff, 1993; Murphy, 2002). This is in line with 8 of the 25 main categories considered in the WordNet hierarchy for nouns denoting tangible things: {*animal*, *fauna*}, {*artifact*}, {*body*}, {*food*}, {*natural object*}, {*person*, *human being*}, {*plant*, *flora*}, {*substance*}. From our point of view these categories can be collapsed in artifactual and natural kinds.

3. Space in language and cognition

Levinson (2004) points out that the spatial thinking is a crucial feature in our lives: we constantly consult our spatial memories in events such as finding our way across town, giving route directions, searching for lost keys, and so on. This importance is mirrored in real discourse where knowledge about formal, agentive, constitutive and telic features, as well as spatial features, are found in specialized domains.

There are three frames of reference lexicalized in language: intrinsic, relative and absolute frame. *Intrinsic frame* involves an object-centred coordinate system, where the coordinates are determined by the "inherent features", sidedness or facets of the objet to be used as the ground (i.e., *he's in front of the house*). *Relative frame* of reference presupposes a viewpoint where a perceiver is located, a figure and ground distinct from the viewpoint. Thus, it offers a triangulation of three points, and utilizes coordinates fixed on viewpoint to assign directions to figure and ground (i.e., *the ball is to the left of the tree*). Finally, *absolute frame* refers to the fixed direction provided by gravity (i.e., *he's north of the house*).

3.1. Work related

Mani *et al*. (2010) focused on the problem of extracting information about places, considering both absolute and relative references. Their goal was on grounding such references to precise positions that can be characterized in terms of geo-coordinates. These authors use a supervised approach to mark up PLACE tags in documents. SpatialML is an annotation scheme derived from this work and which has been applied to annotated corpora in English and Mandarin Chinese. An automatic tagger for SpatialML extents scores 86.9 F-measure, which is a reasonable performance. On the other hand, Clementini et al. (1997) propose a unified framework for the qualitative representation of positional information in a two-dimensional space in order to perform spatial reasoning. The orientation and distance relations for objects modeled as points can determine positional information. The implicit characteristics of an object are its topology and its extension, while, with respect to other objects, topological, orientation, and distance relations have to be considered.

3.2. Axial properties

Evans (2007) explains that a spatial scene is a linguistic unit containing information based on our spatial experience. This space is structured according to four parameters: a figure (or *trajector*), a referent object (that is, a *landmark*), a region and —in certain cases— a secondary reference object. These two reference objects configure a reference frame. We can understand this configuration by considering the following example: *un auto está estacionado detrás de la escuela* (Eng.: "a car is parked behind the school"). In this sentence, *un auto* is the figure and *la escuela* is the referent object. The region is established by the combination of the adverb *detrás*[1] which sketches a spatial relation with the referent object. This relation encodes the location of the figure.

Moreover, Evans (2007) points out the existence of axial properties, that is, a set of spatial features associated to a specific referent object. Considering again the sentence *a car is parked near to the school*, we can identify the location of the car searching for it in the region near to the school. Therefore, this search can be performed because the referent object (the school) has a set of axial divisions: front, back and *side* areas.

3.3. Axial properties and place adverbs

Axial properties are linguistically represented by place adverbs. In this experiment we only consider adverbs functioning in Spanish with preposition *de* (Acosta and Aguilar, 2015):

> *Enfrente*, *delante* (Engl. *In front to/of*); *Detrás*, *atrás* (Engl. *Behind*); *sobre*, *encima* (Engl. *On*); *abajo*, *debajo* (Engl. *under*); *dentro*, *adentro* (Engl. *In/inside*); *fuera*, *afuera* (Engl. *Out/outside*); *arriba* (Engl. *Above/over*).

Additionally, we use some synonymous nouns such as exterior (outside) and interior (in), as well as *side* nouns synonymous with the dimensions left and right.

4. Nominalization

According to Martin (1993: 203-220) and Vivanco (2006), from a linguistic perspective, the discourse neutrality in science and technology is presented by means of im-

[1] In English, *behind* is a preposition. In contrast, in Spanish is an adverb.

personation: missing second person, low presence of first person, abundance of impersonal verbs and passive voice, as well as nominalizations hiding actions made by the subject. These nominalizations are used by scientists to support their arguments, coining new terms by means of nouns and summarizing information previously provided in a text.

In line with the frequent use of nominalization in specialized texts, in the case of Spanish, Cademártori, Parodi and Venegas (2006) show data concerning the use of deverbal nominalizations in three domains: commercial, maritime and industrial. The most used suffixes for constructing nouns are: *-ción*, *-miento*, *-sión*, and *-dor*.

5. Adjectives-Noun modifiers

An adjective is a grammatical category whose function is to modify nouns (Demonte, 1999). There are two kinds of adjectives: descriptive and relational adjectives. The descriptive adjectives refer to constitutive features of the modified noun characterized by means of a single physical property: color, form, character, predisposition, sound, and so on, e.g., *el libro azul* (Eng.: "the blue book"). On the other hand, relational adjectives assign a set of properties, i.e., all the characteristics jointly defining names as *sea*: *puerto marítimo* (Eng.: "maritime port"). In terminology, relational adjectives represent an important element for building specialized terms. For example, *inguinal hernia*, *venereal disease* and others are considered terms in medicine as opposed to NPs with more contextual interpretations like *rare hernia*, *serious disease*, and *critical disorder*.

5.1. Identifying syntactically non-relevant adjectives

If we consider the internal structure of adjectives, we can identify two types: permanent and episodic adjectives (Demonte, 1999). The first kind of adjectives represents stable situations, permanent properties characterizing individuals. These adjectives are located outside of any spatial or temporal restriction (i.e., *psicópata*/"psychopath"). On the other hand, episodic adjectives refer to transient situations or properties implying change and with time-space limitations.

Almost all descriptive adjectives derived of participles belong to this latter class as well all adjectival participles (i.e., *harto*/"jaded"). Spanish is one of the few languages that in its syntax represent this difference in the meaning of adjectives. In many languages this difference is only recognizable through interpretation. In Spanish, individual properties can be predicated with the verb *ser*, and episodic properties with the verb *estar*, which is an essential test to recognize what class an adjective belongs to. In this sense, with the goal of identifying and extracting non-relevant adjectives, we propose extracting adjectives predicated with the verb *estar* (Acosta, Aguilar and Sierra, 2013).

Another linguistic heuristic for identifying descriptive adjectives is that only these kinds of adjectives accept degree adverbs or are part of comparative constructions, e.g., *muy alto*/"very high", *Juan es más alto que Pedro*/"John is taller than Peter".

Finally, only descriptive adjectives can precede a noun because —in Spanish— relational adjectives are always postposed (e.g., *la antigua casa*/“the old house”).

5.2. Types of relational adjectives

According to Bosque (1993) relational adjectives such as *salivary* in the noun phrase *salivary gland* belong to a kind of relational adjectives which do not occupy positions in the argument structure of the predicate, but they denote entities which establish a specific relation with the head noun. Bosque refers to these relational adjectives as *classification relational adjectives*, while the term *thematic relational adjectives* is left for the other group, e.g., the case of *renal infection*, where *infection* is derived from a verb.

6. Methodology

In this paper we propose a methodology for extracting concrete entities from a specialized domain corpus with part-of-speech tags.

6.1. Part-of-Speech Tagging

Part-of-Speech (POS) tagging is the process of assigning a grammatical category to each word in a corpus. The most common taggers used for Spanish are *TreeTagger* (Schmid, 1994) and *FreeLing*[2] (Carreras et al., 2004). In this experiment, we use FreeLing because it is more precise than TreeTagger for tagging texts in Spanish. The following example shows a sentence in Spanish tagged with the FreeLing tagger:

el/DA tipo/NC más/RG común/AQ de/SP lesión/NC ocurrir/VM cuando/CS algo/PI irritar/VM el/DA superficie/NC externo/AQ del/PDEL ojo/NC

6.2. Chunking

Chunking is the process of identifying and classifying segments of a sentence by grouping the major parts-of-speech that form basic non-recursive phrases.

In this work, we concern the automated extraction of concrete entities. Concrete entities relevant to a domain are terms and the most productive patterns of terms consist of a noun and zero or more adjectives (Vivaldi, 2001). Using FreeLing tags, these patterns can be represented as a regular expression in a single pattern:

<NC><AQ>*

The above regular expression is considered in the first phase of extraction of candidates.

[2] FreeLing based on the tags of the EAGLES group.

Concrete entities can be located in spatial scenes as figures or reference objects. In this experiment, only reference objects are extracted with their axial properties that can be linguistically represented as:

<RG|NC><PDEL><DA>?<NC><AQ>*

The regular expressions used to extract non-relevant adjectives according to the linguistic heuristics mentioned in section 5.1 are:

<RG><AQ>
<VAE><AQ>
< D.*|P.*|F.* |S.*><AQ><NOUN>

Where RG, AQ and VAE as tagged with FreeLing, correspond to adverbs, adjectives and the verb *estar*, respectively. Tags <D.*|P.*|F.*|S.*> correspond to determinants, pronouns, punctuation signs and prepositions. The expression <D.*|P.*|F.*|S.*> is a restriction to reduce noise, since elements wrongly tagged by FreeLing as adjectives are extracted without this restriction.

6.3. Bootstrapping phase

We use the candidates to concrete entities obtained in the first step as seeds for extracting more candidates. On the one hand, we assume that coordinating phrases where a good candidate occurs have a high probability of containing other good candidates for a concrete entity:

<NC><AQ>*<CC><NC><AQ>*

Where <CC> tag corresponds to the disjunction (i.e.: *kidney* ***or*** *liver*) and conjunction (i.e.: *kidney* ***and*** *liver*).

On the other hand, noun phrases with at least an adjective take advantage of the noun head of candidates for a concrete entity for finding more specific candidates (i.e., artery-femoral artery):

<NC><AQ>+

6.4. Reducing noise

We sought to remove non-relevant words from noun phrases before ranking candidates for concrete entities. After the chunking phase, noise was reduced by removing non-relevant open-class words. One of our goals consists of building this stopword list as automatically as possible.

Since concrete entities are terms in the domain, a list of non-relevant words from the domain (i.e., stopword list) can be used to refine the terminology obtained from an automatic process. We considered a list constructed with high frequency words in a reference corpus to have drawbacks because, apart from the selection by occurrence frequency (in the domain corpus, words with high frequency can be terms), human supervision is required in order to determine whether a word is relevant to the domain.

Given the above, we consider that linguistic heuristics operating in a specific language can be taken into account in order to automate the selection of non-relevant words. One of the disadvantages, however, is that this leads to language dependence. For the case of adjectives, in Spanish, characteristic features have been proposed in order to distinguish between descriptive and relational adjectives as mentioned in section 5. On the other hand, with a corpus comparison approach, we obtain both nouns and adjectives where the relative frequency in a reference corpus is greater or equal than in the domain corpus. These words can be used as part of the stopword list. Additionally, we take into account empirical evidence concerning the use of deverbal nominalizations in specialized discourse (Cadermártori, Parodi and Venegas, 2006) for removing phrases where noun heads are indicative of actions, events and states but not concrete entities (in a NP with a noun head of this type, a thematic relational adjective is found). In this sense, suffixes as *–ción, -miento*, and *–sión* were used for filtering out noun phrases. Finally, a short list with the more frequent non-relevant nouns operating as noun heads in phrases: *form, type, kind, cause, effect* and so on, were considered for removing noun phrases.

Adjectives from the reference corpus can be used as a fixed-size list where non-relevant adjectives automatically extracted from the domain can be added. These can be obtained taking into account the three heuristics mentioned in section 5.1. Then, these adjectives can be manually reviewed in order to determine their relevance to any specialized knowledge domain (i.e., adjectives as relevant, important, necessary, appropriate, and so on can be considered for the stopword list). This is a fixed-size list and can be the base-list where non-relevant adjectives automatically extracted from the domain can be added.

6.5. Ranking words

We evaluate termhood of simple words by means of rank difference (Kit and Liu, 2008) between two different corpora as in the formula (1). Given the syntactical pattern used for terms in this study, we take into account only nouns and adjectives in both corpora because they are the kind of words most used for building terms:

(1)

Where f_{dom} and N_{dom} correspond to the absolute occurrence frequency of w_i and the size of the domain corpus, respectively. Similarly, f_{ref} and N_{ref} correspond to absolute occurrence frequency of w_i and the size of the reference corpus.

Kit and Liu (2008) only focus on extracting single-word term candidates, so they only weigh words occurring in both the domain and the general corpus. In our experiment we also consider words that only occur in the domain corpus. We assumed that the reference corpus is large enough to filter out non-relevant words, hence words only occurring in the domain corpus have a higher probability of being relevant and the word's frequency reflects its importance:

(2)

We consider that the larger the reference corpus, the higher the *exhaustivity*[3] of open class words of general usage, as well as a higher probability that specialty terms occur at least one time (the reference corpus was collected from an online newspaper where news about science and technology are published too), so that we would expect a higher precision in ranking.

6.6. Ranking multi-word term candidates

Formally, if a candidate noun phrase (np) has a length of n words, $w_1 w_2 \ldots w_n$, where $n>1$, then the ranking of the candidate *np* is the sum of the frequency of *np* as a whole plus the weights of all the individual words w_i:

(3)

7. Results

This section presents the results of our experiment considering a subset of 1,200,000 tokens of the MedLineplus corpus.

7.1. Sources of textual information

Domain corpus

The source of textual information is constituted by a set of documents of the medical domain, basically human body diseases and related topics (surgeries, treatments, and so on). These documents were collected from MedlinePlus in Spanish.

The size of the corpus is 1.2 million tokens, but we carried out our experiment with a subset of 200,000 words in order to determine manually the number of concrete entities present in the results. As an ongoing work, we are manually determining how many concrete entities are present in the complete corpus. We chose a medical domain due to the availability of textual resources in digital format. Finally, we assume that the choice of domain does not suppose a very strong constraint for generalizing the results to other domains.

Reference corpus

With the goal of ranking words relevant to the domain by means of their relative frequency ratio, a large reference corpus was collected from an online newspaper[4] with new articles from 2014 (the size of corpus is about 5 million tokens). URLs from the

---

[3] *Exhaustivity* of a document description is the coverage it provides for the main topics of the document. So, if we add new vocabulary terms to a document, the *exhaustivity* of the document description increases (Baeza and Ribeiro, 2011).

[4] www.lajornada.com.mx. Mexican newspaper with information available online.

main heads were automatically extracted using the Python library BeautifulSoup[5]. Then, this set of URLs was introduced in WebBootCat, a search tool of Sketch Engine[6], in order to automatically collect the textual information from each WEB page. The description of the structure of the reference corpus is showed in table 1.

Table 1. Structure of the reference corpus.

Category	**Docs**	**%**
Sciences	24	0.4
Politics	1865	29.3
Entertainment	98	1.5
Sports	515	8.1
Society	416	6.5
City	424	6.7
States	449	7.1
Economy	658	10.4
World	662	10.4
Culture	137	2.2
Editorial	316	5.0
Mails	318	5.0
Opinion	319	5.0
Homepage	155	2.4

7.2. Other resources

The programming language used in order to automate all tasks required was Python version 3.4 as well as the *NLTK* module version 3.0 (Bird, Klein and Loper, 2009). Additionally, the POS tagger used in this experiment was *FreeLing* which is included in *Sketch Engine*.

[5] www.crummy.com/software/BeautifulSoup/bs4/doc/

[6] https://the.sketchengine.co.uk

7.3. Analysis of results

The first phase of extraction of candidates to concrete entity without filters achieves a global precision of 56%. The tables 2 and 3 show precision with different thresholds of candidates starting with the better ranked candidates. With the stopword list built as mentioned in section 6.4, we achieve a global precision of 76%. Global precision with a stopword list reflects an improvement of 20%, but a significant loss of 17% of true candidates. As can be seen from these tables, the ranking of words and noun phrases is useful for sorting results from the most relevant to the least relevant results.

Table 2. Comparison of results.

Candidates	**Precision**	
	Without filter	**With filter**
100	91%	96%
200	87%	87%
300	73%	83%
400	69%	
500	63%	

Bootstrapping phase

The bootstrapping phase taking into account coordinating phrases achieves a set of 1248 candidates, of which 262 are new true candidates. The global precision with this second phase is of 47%, with a precision by thresholds as shown in table 3. The advantage of this phrase structure is that single-word candidates can be extracted.

On the other hand, the bootstrapping phase considering noun phrases achieves a set of 2796 candidates, of which 1534 are good candidates. The global precision of this phase is of 55%, with a precision by thresholds as shown is table 3. One disadvantage of this structure is that only candidates with at least one adjective can be selected.

Table 3 shows a better performance with noun phases. The identification of the concrete entities present in corpus is an ongoing task that will let us evaluate in terms of recall too.

Table 3. Bootstrapping phase.

Candidates	Coordinating phrases	Noun phrases
100	55%	71%
200	59%	71%
300	59%	69%
400	59%	68%
500+	53%	65%

7.4. Discussion

The candidates in a bootstrapping phase give us insight about the kind of semantic relations implicit in noun phrases of the type <*NC*><*AQ*>. Given the phase of reduction of non-relevant adjectives, we have a great deal of relational adjectives where it is possible to find different relations. For example, *salivary gland* has implicit a telic relation. On the other hand, *testicular gland* has a part-whole or locative relation. Finally*, meibomian gland* may be considered as a specific type of *gland.*

With respect to the extraction of lexical relations, specifically hyponymy-hypernymy relations (Hearst, 1992; Wilks, Slator and Guthrie, 1995; Pantel and Pennacchiotti, 2006), as well as meronymy relations (Berland and Charniak, 1999; Girju, Badulescu and Moldovan, 2006), these works are based on patterns where two terms are located in the context of a sentence: the hand has fingers, the dog is an animal, and so on, but there are few jobs working with noun phrases, which we consider it is very important because we could consider a noun phrase as salivary gland as an hyponym of gland, but it is clear that if we dig a little deeper that the semantic relation implicit is telic.

8. Conclusions

We discussed a methodology for extracting concrete entities in the medical domain. Concrete entities have been studied since Aristotle's works, particularly in his biological and zoological descriptions. According to Aristotle's categories (the first category), many things can be predicated of substances. We assume that substances are concrete entities, with a more extended meaning, i.e.: the eight tangible categories formulated by Fellbaum for WordNet (1998). Thus, we consider that the automated identification and extraction of this kind of information is an important advance in further NLP tasks.

Cognitive abilities as the spatial knowledge and his representation in natural language are important for our extraction methodology. We observe that spatial descriptions are frequent in specialized discourses. Additionally, we propose a further step of bootstrapping in order to find a great number of candidates for concrete entities. Candidates with a concrete entity as a noun head and a relational adjective show semantic relations as part-whole, locative, agentive and telic, which can be interpreted, at first, as hyponymy/hyperonymy relations.

On the other hand, to assign relevance to words is an important step for ranking candidates, according to our exposed results. In this sense, as ongoing work, we are collecting more information about science and technology at the same electronic journal in order to improve the results in the ranking process.

Finally, it is necessary to mention that POST taggers as FreeLing and TreeTagger fail in the task of identifying nouns, adjectives and verbs closely related with the domain. This failure has a negative impact on the results. We believe it is important to face this problem in future extraction tasks.

Acknowledgments

This paper has been supported by the National Commission for Scientific and Technological Research (CONICYT) of Chile, Project Numbers: 3140332 and 11130565.

9. References

1. Acosta, O., Aguilar, C. & Sierra, G. Using Relational Adjectives for Extracting Hyponyms from Medical Texts. In A. Lieto & M. Cruciani (eds.), *Proceedings of the First International Workshop on Artificial Intelligence and Cognition (AIC 2013)*, CEUR Workshop Proceedings, pp. 33-44.Torino, Italy. (2013).
2. Acosta, O. & Aguilar, C. Extraction of Concrete Entities and Part-Whole Relations. In B. Sharp & R. Delmonte (eds.), *Natural Language Processing and Cognitive Science. Proceedings 2014*, pp. 89-100. Berlin, De Gruyter (2015).
3. Baeza, R. & Riveira, B. *Modern Information Retrieval*, 2nd ed. New York, Addison Wesley (2011).
4. 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. College Park, Maryland, USA, ACL Publications (1999).
5. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python*, Sebastropol, Cal., O'Reilly (2009).
6. Bosque, I. Sobre las diferencias entre los adjetivos relacionales y los calificativos. *Revista Argentina de Lingüística*, No. 9, pp. 10-48 (1993).
7. Carreras, X. Chao, I., Padró, L. & Padró, M. FreeLing: An Open-Source Suite of Language Analyzers. In M.T. Lino *et al.* (eds.) *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004*, pp. 239-242. Lisbon, Portugal, ELRA Publications (2004).
8. Cademártori, Y., Parodi, G. & Venegas, R. El discurso escrito y especializado: caracterización y funciones de las nominalizaciones en los manuales técnicos, *Literatura y Lingüística*, No. 17, pp. 243-265 (2006).
9. Chunyu, K. & Liu, X. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2), 204-229 (2008).

10.Clementini, E., Di Felice, P., & Hernández, D. Qualitative representation of positional information. *Artificial intelligence*, *95*(2), 317-356 (1997).
11.Demonte, V. El adjetivo. Clases y usos. La posición del adjetivo en el sintagma nominal. In I. Bosque & V. Demonte (eds.), *Gramática descriptiva de la lengua española*, Vol. 1, Cap. 3, pp. 129-215. Madrid, Espasa-Calpe (1999).
12.Evans, V. *A Glossary of Cognitive Linguistics*, Edinburgh, UK, Edinburgh University Press (2007).
13.Fábregas, A. The internal syntactic structure of relational adjectives, *Probus*, 19(1), 1-36 (2007).
14.Fellbaum, C. *WordNet: An Electronic Lexical Database*, Cambridge, Mass., MIT Press (1998).
15.Girju, R., Badulescu, A. & Moldovan, D. Automatic discovery of part–whole relations. *Computational Linguistics*, 32(1), 83-135 (2006).
16.Hearst, M. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 539-545, Nantes, France. ACL Publications (1992).
17.Landau, B. & Jackendoff, R. What and where in spatial language and spatial cognition, *Behavioral and brain sciences*, 16(02), 255-265 (1993).
18.Levinson, S. *Space in Language and Cognition: Explorations in Cognitive Diversity*, Cambridge, UK, Cambridge University Press (2004).
19.Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J. & Clancy, S. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3), 263-280 (2010).
20.Martin, James R. Technicality and abstraction: Language for the creation of specialized texts. In M.A.K. Halliday & James R. Martin. *Writing science: Literacy and discursive power*, pp. 203-220, London, The Falmer Press (1993).
21.Murphy, G. *The Big Book of Concepts*. Cambridge, Mass., MIT Press (2002).
22.Pustejovsky. J. *The generative lexicon*, Cambridge, Mass., MIT Press (1996).
23.Schmid, H. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Vol. 12, pp. 44-49. Manchester, UK (1994).
24.Smith, B., and Kumar, A. Controlled vocabularies in bioinformatics: a case study in the gene ontology, *Drug Discovery Today: BIOSILICO*, 2(6), 246-252 (2004).
25.Vivanco, V. *El español de la ciencia y la tecnología*, Madrid, Arco Libros (2006).
26.Vivaldi, J. *Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas*. PhD Dissertation. Barcelona, Universidad Politècnica de Catalunya (2001).
27.Wilks, Y., Slator, B. & Guthrie, L. *Electric Words,* Cambridge, Mass., MIT Press (1995).
28.Winston, M., Chaffin, R. & Herrmann, D. A taxonomy of part−whole relations, *Cognitive science* 11(4), 417-444 (1987).