Definitional verbal patterns for semantic relation extraction

Gerardo Sierra, Rodrigo Alarcón, César Aguilar and Carme Bach

In this paper we present a description of the role of definitional verbal patterns for the extraction of semantic relations. Several studies show that semantic relations can be extracted from analytic definitions contained in machine-readable dictionaries (MRDs). In addition, definitions found in specialised texts are a good starting point to search for different types of definitions where other semantic relations occur. The extraction of definitional knowledge from specialised corpora represents another interesting approach for the extraction of semantic relations. Here, we present a descriptive analysis of definitional verbal patterns in Spanish and the first steps towards the development of a system for the automatic extraction of definitional knowledge.

Keywords: Definitional context, definitional verbal patterns, definitional knowledge extraction.

1. Introduction

The possibility of searching for and recognising semantic relations in definitions occurring in specialised text corpora is one of the current applications of computational lexicography and terminology. A detailed linguistic analysis of the relationship established between a definition and all the elements that permit its insertion in a discursive context, specifically, those verbal patterns whose function is to associate a definition to its corresponding term is thus necessary.

In this paper we therefore propose the design of an automatic extractor of definitional knowledge based on a set of semantic relationships, taking into account that these types of relationships have syntactic representations in constructions that we have named *definitional verbal patterns.*

As a starting point, we will show how the extraction of semantic relations based on lexicographic definitions (mainly genus and differentia) follows the model IS-A. However, textual corpora offer other types of definitions from which we can extract semantic relations that go beyond the analytical pattern. Therefore, we will also show how we can use a discursive structure named *definitional context* (DC) for the recognition and extraction of definitions from specialised texts using regular verbal patterns. Consequently, we propose a grammatical analysis of these verbal patterns following a grammatical formal model, such as the Government and Binding Theory. We will then formulate a possible typology of definitions based on the semantic relationship that each definition establishes with a specific verbal pattern. With these data, and bearing in mind the relationships introduced by the analysed verbal patterns, we will describe a set of experiments to extract definitions. These experiments constitute an empirical support for the design of a system for the automatic extraction of definitional contexts. Such a system includes the extraction of each occurrence of a definitional pattern, the filtering of non-relevant contexts and the identification of DC's constitutive elements, i.e., terms, verbal patterns and definitions. This system is being developed for Spanish. The evaluation of its performance is carried out using the Precision & Recall method. Finally, we describe the future work.

2. The role of definitional contexts for the extraction of semantic relations

2.1 Finding semantic relations in definitions

Recent work in computational linguistics has shown machine readable dictionaries (MRDs) to be a promising lexical information source for development of taxonomies and automatic extraction of certain semantic relations. There is a general agreement that dictionaries are good repositories of lexical and taxonomic information (Calzolari and Picchi 1988; Boguraev and Pustejovsky 1996) and that this information can be extracted using computational techniques through the analysis of the definition of an entry (Pustejovsky et al. 1993).

One of the most important semantic relationships to be extracted from MRDs for clustering relies on the identification of two well known data, namely *genus term* and *differentia*. The former is a more general term than the headword while the later is the set of words that serves to differentiate the headword from other headwords with the same Genus. The earliest identification of the usefulness of the genus term and differentia for taxonomy construction was presented by Amsler (1981). Using the *Longman Dictionary of Contemporary English* (LODCE), Alsawhi (1987) proposed a method for extracting and categorising lexical definitions based on the phrasal operator IS-A. He identified the *semantic head* (that is, the genus term of an analytical definition), and other information occurring in definitions, such as properties, purpose and predications.

As a part of a multilingual research project named *Aquilex*, Vossen and Copestake (1993) use again the semantic operator IS-A to delineate a set of taxonomies for classifying lexical definitions. Basically, they propose 3 types of semantic relationships:

- **Hyponymy-Hyperonymy:** a hyponymic entity is derived from a superior hyperonymous, for example *an autobiography IS-A book.*
- **Synonymy:** two entities that maintain certain equivalence at the cognitive level (Cruse 1986), for example: *a policewoman IS-A female policeman.*
- **Individuation:** those entities where a shift of individuation takes place. There are two kinds of individuation: a) *quantity/mass*, i.e., a relationship between a portion or a piece and a certain substance or entity, e.g. *an hour IS-A portion of time*; b) *member/group*, which is a relationship between an entity that can be inherent to a group or collective, e.g. *a policeman IS-A member of a police force*.

2.2 Semantic relations and corpora

As Sager and Ndi-Kimbi (1995) observe, analytical definitions (usually given in dictionaries) are not sufficient for describing every possibility to formulate a definition in natural language. Therefore, it is also necessary to consider definitional forms different from the analytical model. Other means for obtaining semantic relations beyond

dictionaries is the extraction and retrieval of definitions from scientific and technical corpora.

Pearson (1998) identifies three types of definitions based on the degree of proximity or distance to the analytical model:

- **Formal definitions**. This type of definition is closer to the analytical model; it is characterised by the formal scheme *X = Y + Features*, i.e., a term equal to a genus term plus a differentia.
- **Semi-formal definitions**. This type is more frequently found in technical texts and can be represented by the scheme *X = Features*, where there is a term and a differentia without a genus term.
- **Non-formal definitions**. This type of definition does not conform to a specific scheme, since it may have multiple representations using both linguistic (verbal phrases, adjectives, adverbs, etc.) and non-linguistic (for example, typographical marks) structures.

Condamines and Rebeyrolle (2001) show another kind of non-analytical definitions used in discursive contexts. By analysing patterns contained in definitions it is possible to obtain other types of semantic relations. For example, a semantic frame of trajectory in constructions such as: *The Component development cycle takes place during the product realisation phase*, where the phrase *takes place during* supposes a temporal frame that delimits the beginning and the end of an action.

2.3 Definitional knowledge extraction

In accordance with authors such as Morin (1998) or Malaisé et al. (2005), we believe that definitions contained in specialised texts could be considered the departing point for extracting semantic relations. Therefore, the automatic extraction of definitions from DCs is an important step for obtaining semantic relations within the specialised knowledge of any scientific or technical field.

The extraction of definitional knowledge has been approached from both theoretical–descriptive studies and applied research.

One of the first theoretical–descriptive studies was carried out by Pearson (1998). She describes the behaviour of contexts where terms occur. She states that, when authors define a term, they usually use typographical patterns to visually highlight the presence of terms and/or definitions as well as lexical and metalinguistic patterns to connect terms with their definitions by means of syntactic structures.

This idea was reinforced by Meyer (2001), who stated that definitional patterns can also provide keys that allow the identification of the type of definition found in discursive contexts which is a helpful tool in the development of ontologies.

More recently, Feliu (2004) studied the recurrent verbs found in different semantic relations in order to establish a typology for its classification and its further automatic extraction.

The search for discursive markers, which are textual fragments where a reformulation process is given, constitute another means for extracting definitional knowledge. Bach (2005) describes and analyses the role of these markers in the process of extraction of definitional knowledge and states that they should be taken into account at the time when automatic search for relevant information about terms is performed.

Generally speaking, theoretical-descriptive studies share the idea that DC's extraction is possible by searching for recurrent definitional patterns. These patterns can be embodied as typographical and lexical patterns. Typographical patterns refer to text typography or punctuation marks while lexical patterns refer to syntactic patterns e.g. definitional verbs or reformulating markers.

On the other hand, applied investigations aim to elaborate methodologies for the automatic extraction of definitional knowledge taking into account the results of theoretical–descriptive studies. In particular, some of these investigations are focused on:

- The automatic identification of definitions in English medical texts (Klavans and Muresan 2001).
- The extraction of definitions for question answering systems for English (Saggion 2004)
- The extraction of metalinguistic information for English (Rodríguez 2004).

- The identification of relevant information for the elaboration of ontologies for French (Malaisé 2005).
- The extraction and annotation of definitions in a German language corpus (Storrer and Wellinghoff 2006).
- The automatic identification of semantic relations between two specific terms for Catalan (Feliu et al. 2006).

The main purpose of these applied investigations is the extraction of relevant information about terms, mainly in English, French, Catalan and German, but nothing has been done for Spanish. As well as in theoretical-descriptive studies, the specific purpose of each applied investigation is different but they all share some specific concrete ideas.

The first of these ideas is to start by searching for definitional patterns, either typographical or syntactic. Those authors agree that the extraction of occurrences of definitional patterns is a good starting point for finding terms and definitions. However, the process of searching for occurrences of definitional patterns can also extract noise, i.e., non-relevant contexts where definitional patterns occur. This noise could be filtered by studying the cases where definitional patterns occur in a more general sense, and by developing techniques for its automatic filtering.

Applied investigations also state that, once the occurrences of definitional patterns have been extracted, an automatic analysis of these occurrences should be carried out to identify the essential information extracted, mainly terms and definitions. Each author proposes different ways to solve this identification, for example algorithms that process syntactic information (Saggion and Gaizauskas 2004) or processes for the identification of recurrent terms and definition's position based on the pattern that establishes the relation between them (Malaisé et al. 2005).

Following these studies, we consider a Definitional Context (DC) the discursive context where relevant information to define a term could be found (Alarcón and Sierra 2003). The minimal constitutive elements of a DC are: a term, a definition, and usually linguistic or metalinguistic forms as verbal phrases, typographical markers and/or pragmatic patterns (mainly explicit information about how the term should be understood). 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.* (Engl. 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.)

In this case, we can see that the DC sequence is formed by the term *energía primaria* (Engl. primary energy), the definition *aquel recurso…* (Engl. 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* (Engl. in general terms) and the typographical marker (bold font) that in this case emphasises the presence of the term.

3. Definitional verbal patterns related to specific semantic relationships

The linguistic analysis of DCs obtained from specialised texts shows a set of regular verbal patterns in Spanish whose function is to introduce the definitions and to link them with their terms. We call these patterns *definitional verbal patterns* (DVPs), and they link terms and definitions in a kind of syntactic chain. For example, a verb like *Ser* (Engl. to be) constitutes a syntactic structure with a noun phrase (NP) to the left of the verb, a verbal phrase (VP) as a connector, and a predicate to the right of the verb, the later represented by a syntactic phrase such as an adjective phrase (AdjP), adverbial phrase (AdvP), noun phrase (NP), prepositional phrase (PP), inflectional phrase (IP) or concordance phrase (CP). An example of this is:

[[*La cuchilla fusible* $_{NP}$] [***es*** [*un elemento de conexión y desconexión de circuitos eléctricos* $_{VP}$] $_{IP}$]. (Engl. [[The fuse-switch disconnector $_{NP}$] [**is** [an element of connection and disconnection of electric circuits $_{VP}$] $_{IP}$].]

Here, the term *cuchilla fusible* (Engl. fuse-switch disconnector) is just to the left of the verb *es* (Engl. is), and the definition *un elemento de conexión y desconexión.…* (Engl. an element of connection and disconnection…) is the predicate located to the right of the verb *es* (Engl. is). This kind of grammatical structure introduces an analytical definition, where the genus term is represented by a NP *un elemento* (Engl. an element) and the differentia is

represented by a prepositional phrase PP *de conexión y desconexión de circuitos eléctricos* (Engl. of connection and disconnection of electric circuits).

We used a syntactic-formal model, the Predication Theory, to analyse these DVPs. The Predication Theory (or PT) is a model derived from Government & Binding Grammar (or GB) formulated by Chomsky (1981). We believe GB to be a pertinent theoretical framework for describing the syntactic behaviour of DVPs, bearing in mind the following arguments:

- GB is a formal grammar that offers a detailed explanation of predicative patterns, considering that their syntactic and semantic features are described in a simple tree-model. This kind of description is relevant for the automatic parsing of natural language patterns (Karttunen and Zwicky 1985).
- Other types of formal grammars derive from GB, such as Lexical-Functional Grammar, Tree Adjoining Grammar, Head-Driven Phrase Structure Grammar and so on (Sells 1985). We think the description of DVPs within the framework of GB is relevant for both information extraction and term extraction through grammatical formalisms.
- Finally, the use of GB to identify syntactic patterns associated with terms has proved successful in automatic systems like LEXTER (Bourigault et al. 1996) and Syntex (Bourigault et al. 2005). We think the degree of formality of GB is relevant for developing algorithms both to recognise and extract syntactic patterns associated with DVPs.

PT is a formal model derived from GB that offers an appropriate syntactic description of DVPs. *Grosso modo*, PT states that all predications indicate a semantic relationship between an entity and a particular property or characteristic feature (Williams 1980; Napoli 1989; Bowers 2003).

We consider two types of predicative phrases (henceforth, PredP): a) a *simple* or *primary predication*, i.e., those PredP conformed by a subject to the left of the verb, and a predicate that is located to the right of the verb; b) a *double* or *secondary predication*, which integrates a subject in a pre-verbal position, and an object and the predicate, both

after the verb. In the latter case, the predicate can affect either the subject or the object of a sentence. For example:

1. a. **Turing** define **una computadora** *como un mecanismo electrónico que procesa conjuntos de datos.* (Engl. Turing defines a computer as a kind of electronic device that processes a set of data.)
   b. **Turing** define **una computadora** *conforme a su teoría.* (Engl. Turing defines a computer according to his theory.)

In (1a), the predicate *como un mecanismo electrónico…* (Engl. as a kind of electronic device…) affects the object *una computadora* (Engl. a computer). In (1b), the predicate attributes the feature *conforme a su teoría* (Engl. according to his theory) to the subject *Turing*, and not to the object *una computadora* (Engl. a computer). This distinction is explained, for Spanish by Demonte (1987) and Mallén (1991).

Coming back to Sager and Ndi-Kimbi's proposal (1995) regarding the existence of alternative verbal patterns for expressing concepts in natural language, we believe there is a relationship between definitions and a DVP. As a starting point, it is convenient to know the different kinds of definitions. In a previous work, Sierra et al. (2003) established a typology of definitions based on the two basic constituents of the analytical model: Genus and Differentia. The typology depends on whether these constituents are present or absent in the definition (Figure 1).

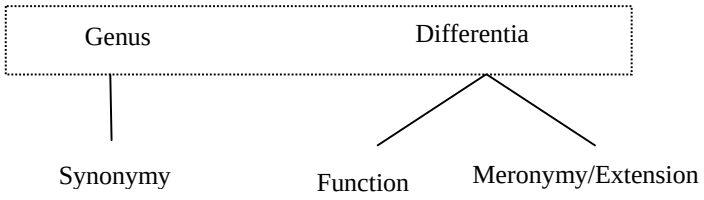


Figure 1. Typology of definitions according to the analytical model

Where:

- Analytical definition: This definition occurs when both constituents, Genus and Differentia, are present.
- Synonymical definition: This definition only provides a genus that is semantically equivalent to the defined term.

- Functional definition: Here the Genus no longer is present but the Differentia indicates the function of the entity.
- Meronymical or extensional definition: This definition has no Genus but includes a differentia that enumerates the parts that make up the entity.

We will briefly describe each type of definition in relation to the associated specific verbal pattern.

3.1 Analytic verbal patterns

3.1.1 Verbal pattern associated with a primary predication.

This pattern configures a structure of the type Subject + Predicate, and is represented by the verbs *ser* (Engl. to be), *representar* (to represent), *referir a* (Engl. to refer to), and *significar* (Engl. to signify/to mean). For example,

2. a. [[El apartarrayos [**es** [un dispositivo [que nos permite proteger las instalaciones contra sobretensiones de origen atmosférico $_{CP}$] $_{NP}$] $_{VP}$] $_{PredP}$]$_{IP}$]. (Engl. [[The lightning conductor [**is** [a device [that allows to protect the electrical systems against surges of atmospheric origin $_{CP}$] $_{NP}$] $_{VP}$] $_{PredP}$]$_{IP}$].)

In this case, the term *El apartarrayos* is to the left of verb *es* (Engl. is) in the position of Subject (Engl. the lightning conductor), and the definition is the Predicate: *un dispositivo que nos permite proteger las instalaciones.*… (Engl. that allows to protect the electrical systems…). The Genus is *un dispositivo*, and the Differentia is *que nos permite*…

3.1.2 Verbal pattern associated with a secondary predication.

Another verbal structure associated with the analytical pattern is the secondary predication Subject + Object + Predicate, where the adverb *como* (in English, the preposition as/like), or the preposition *por* (Engl. for) introduce the Predicate. Here, we group the verbs: *caracterizar* + *como/por* (Engl. to characterise + as/for), *comprender* + *como* (Engl. to comprehend + as), *concebir* + *como* (Engl. to conceive + as), *conocer* + *como* (Engl. to know + as), *considerar* + *como* (Engl. to consider + as), *definir* + *como* (Engl. to define + as), *describir* + *como* (Engl. to define + as), *entender* + *como* (Engl. to understand + as) ,

identificar + *como* (Engl. to identify + as) and *visualizar* + *como* (Engl. to visualize + as). We can represent this pattern in the following way:

2. b. [[*Carlos Godino* $_{NP}$] [***define*** [*la Arquitectura Naval* [*como la ciencia que trata de los conocimientos necesarios para la construcción de los buques* $_{PredP}$] $_{NP}$] $_{VP}$] $_{IP}$]. (Engl. [[Carlos Godino $_{NP}$] [***defines*** [*the Naval Architecture* [*as the science dealing with the necessary knowledge for the construction of ships* $_{PredP}$] $_{NP}$] $_{VP}$] $_{IP}$].)

In semantic terms, the NP *Carlos Godino* is the Agent or Actor (+/- Animate, +/- Human) that performs the act to define a term. The NP *la Arquitectura Naval* (Engl. the Naval Architecture) is equivalent to the term, and the Predicate is the analytical definition formed by the Genus *la ciencia* (Engl. the science) and the Differentia *que trata de los conocimientos necesarios* … (Engl. dealing with the necessary knowledge…).

3.2 Functional verbal patterns

The functional verbal pattern introduces a type of definition where the genus is absent, but introduces a differentia that, semantically, describes the function or the use of a particular entity. The verbal pattern is also associated with a primary predication, therefore, the term is equivalent to the NP to the left of the verb and the definition is the Predicate. The verbs associated with these patterns are: *emplear* (Engl. to employ), *encargarse* (Engl. to be in charge of), *funcionar* (Engl. to function), *ocuparse* (Engl. to occupy/to have the function of), *permitir* (Engl. to permit), *servir* (Engl. to serve), *usar* (Engl. to use) and *utilizar* (Engl. to utilise). Prepositions are sometimes used to link the verb to the pattern: *de* (Engl. of) and *para* (Engl. for). The syntactic representation is:

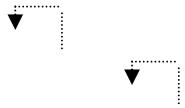
2. c. [[La técnica de velocimetría de imágenes de partícula $_{NP}$], [**permite** [medir la velocidad de un campo de flujo bi o tri dimensional $_{VP}$] $_{PredP}$] $_{IP}$]. (Engl. [[*The method of particle image velocimetry* $_{NP}$] [***allows*** [to *measure the speed of a fluid field in two or three dimensions* $_{VP}$] $_{PredP}$] $_{IP}$].

The term is the NP *La técnica de velocimetría de imágenes de partícula* with no Genus term, and the Differentia describes the function by the VP *medir la velocidad de un campo…* The verb *permite* (allows) corresponds to the PVD.

3.3 Meronymic and extensional verbal patterns

This pattern establishes a type of definition that enumerates the parts or components of a whole (Vossen and Copestake 1993). In a similar way to the previous ones, this pattern is structured around a primary predication. The verbs associated with these patterns are: *componer* (Engl. to compose), *comprender* (Engl. to include), *consistir* (Engl. to consist), *constar* (Engl. to constitute), *contar* (Engl. to count), *constituir* (Engl. to constitute), *incluir* (Engl. to include), and *integrar* (Engl. to integrate). Two specific prepositions can be linked to these verbs: *de* (Engl. of) and *con* (Engl. with):

2. d. [[La zona límite $_{NP}$] [**incluye** $_{t}$ [$_{t}$ planicies costeras, marismas, áreas de inundación, playas, duna y corales $_{NP}$] $_{VP}$] $_{PredP}$]



Engl. [[*The border zone* $_{NP}$] [***includes*** $_{t}$ [$_{t}$ *coastal plains, salt marshes, flood areas, beaches, dunes and corals* $_{NP}$] $_{VP}$] $_{PredP}$]

In this case, the verb *incluye* (Engl. includes) expresses no Genus but a kind of partition that shows several components of the term *zona límite* (Engl. border zone): *planicies costeras* (Engl. coastal plains), *marismas* (Engl. salt marshes), *áreas de inundación* (Engl. flood areas) and so on.

3.4 Synonymic verbal patterns

A synonymic verbal pattern formulates a type of equivalence between a term and the definition, specifically with the genus but not with the differentia. The associated verbs for this pattern are: *equivaler* (Engl. to equivalent), *llamarse* (Engl. to call), *nombrarse* (Engl. to name), *ser* + *igual* (Engl. be + equal), and *ser* + *similar* (Engl. be + similar). In some cases, these verbs can introduce the prepositions *a* (Engl. to), the adverb *también* (Engl. also), and the prepositional phrases *igual a* (Engl. equal to) or *similar a* (Engl. similar to):

2. e. [[la tensión de base [se le **llama también** [tensión unidad $_{NP}$] $_{VP}$] $_{PredP}$]$_{IP}$]. (Engl. [[the base tension [it

is **also called** [unit tension $_{NP}$] $_{VP}$] $_{PredP}$]$_{IP}$].)

In the example, the NPs *la tensión de base* (Engl. the base tension) and *tensión unidad* (Engl. unit tension) show a kind of cognitive equivalent, being the later a Genus of the former.

In Table 1, we summarise the typology we have established and the relationship with the PVD.

Table 1. Verbs associated with definitions

Definition	**Verbs**	**Associated words**	**Predication**
Analytical	*referir* (to refer to) *representar* (to represent) *ser* (to be) *significar* (to signify/to mean)	*a* = to (preposition) (in the case of *referir*, it is a phrasal verb that inserts obligatory the preposition *a*)	Primary predication
Analytical	*caracterizar* (to characterise) *comprender* (to comprehend) *concebir* (to conceive) *conocer* (to know) *considerar* (to consider) *definir* (to define) *describir* (to describe) *entender* (to understand) *identificar* (to identify) *visualizar* (to visualise)	*como* = as/like (adverb) *por* = for/by (preposition)	Secondary predication
Functional	*emplearse* (to employ + clicit "se") *encargar* (to be in charge of) *funcionar* (to function) *ocupar* (to occupy) *permitir* (to permit) *servir* (to serve) *usar* (to use) *utilizar* (to utilise)	*de* = of (preposition) *para* = for (preposition)	Primary predication

Meronymy/ Extensional	*componer* (to compound) *comprender* (to include) *consistir* (to consist) *constar* (to consist) *contar* (to count) *constituir* (to constitute) *contener* (to content) incluir (to include) integrar (to integrate)	*de* = of (preposition) *por* = for/by (preposition) *con* = with (preposition)	Primary predication
Synonymy	*equivaler* (to be equivalent to) *llamar* (to call) *nombrar* (to name) *ser _ igual* (to be equal to) *ser _ similar* (to be similar to)	*también* = also (adverb) a = to (preposition) *igual a* = equal to (adverb phrase) *similar a* = similar to (adverb phrase)	Primary predication

3.5 Structural classification

Furthermore, we have adapted this description of patterns in order to extract them automatically, according to a structural criterion. Depending on whether the pattern includes or not a grammatical particle (e.g., preposition, adverb, adjective, etc.), we classify them in two groups:

- **Single definitional verbal pattern (SDVP).** This group includes patterns which are formed by a single verb presented in a simple form, without any other grammatical particle: X *significa* (Engl. signifies) Y, Y *denominado* (Engl. denominated) X.
- **Compound definitional verbal patterns (CDVP)**. This group is formed by structures that include a verb plus a grammatical particle such as adverbs or prepositions. For example: X *se define como* (Engl. is defined as) Y, X *consiste de* (Engl. consists of) Y.

In the following lines we analyse in detail the use and the probability of finding a specific type of definition using a specific DVP, taking into account the syntactic and semantic relationships that both elements establish in a predicative structure. We also propose the

development of an automatic method for the extraction of DCs that includes definitions with their particular DVPs.

4. ECODE, a definitional context extraction system

In this section we will explain the process for developing a DCs automatic extractor based on the typology established in section 3. Our starting point is the automatic search for and recognition of DVPs. The system includes three modules related to the extraction of DVPs occurrences, the filtering of non-relevant contexts, and the identification of constitutive elements: term, definition and DVP.

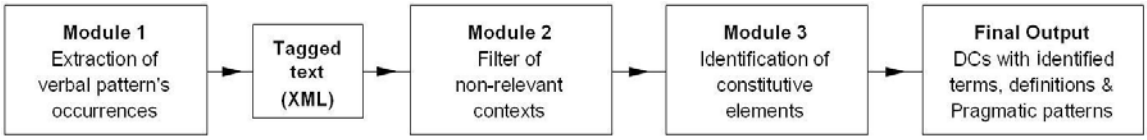


Figure 2. ECODE architecture

We can observe the general architecture of the system in figure 2. The first module extracts occurrences of DVPs from a tagged corpus. These occurrences are also tagged using an XML annotation scheme which is helpful in the next modules. Then, module 2 filters non-relevant contexts, i.e. contexts that do not provide definitional information. Finally, the third module identifies terms, definitions and pragmatic patterns. The final output is a list of DCs with its constitutive elements highlighted.

In the next subsections we will describe the corpus we used for this purpose. Then we will describe each module and show the results obtained as well as an evaluation of them.

4.1 Corpus

We took the IULA´s Technical Corpus and its search engine BwanaNet[1] as a reference. This corpus was developed by the Institut Universitari de Lingüística Aplicada, Universitat

Pompeu Fabra. It consists of documents belonging to the specialised fields of Law, Human Genome, Economy, Environment, Medicine, Informatics and General Language. Up to July 2006 it had a total of 1,011 documents.

BwanaNet, on the other hand, is the search engine of this corpus, which allows users to search for frequencies and simple, standard or complex concordances.

4.2 Extraction of occurrences of definitional verbal patterns

For the experiments described in this paper, we decided to work with a set of verbal patterns representing the divergence of the different definition types mentioned in section 3. We chose at least two verbal patterns for each definition type (Table 2). In the case of analytical definitions, we also studied the behaviour of verbs which can be used in a wider range of contexts, not necessarily definitional contexts, like *concebir* (Engl. to conceive) or *identificar* (Engl. to identify).

Table 2. DVPs employed for definitional knowledge extraction

Verbal pattern		**Definition Type**
Concebir (como)	to conceive (as)	analytical
definir (como)	to define (as)	analytical
entender (como)	to understand (as)	analytical
identificar (como)	to identify (as)	analytical
significar	to signify	analytical
Consistir en	to consist in	extensional
Consistir de	to consist of	extensional
constar de	to comprise	extensional
usar como / para	to use as / for	functional
utilizar como / para	to utilise as / for	functional
servir para	to serve for	functional
Denominar también	also denominated	synonymic
llamar también	also called	synonymic

We searched for each pattern in the IULA's Technical corpus through BwanaNet's complex search option, which allows users to obtain the occurrences with Part-of-Speech (POS) tags. The search was limited to no more than 300 occurrences for each verbal pattern, using the random recovery option. The average amount of retrieved occurrences was around 250 for *definir* (Engl. to define), *entender* (Engl. to understand), *identificar* (Engl. to identify), *consistir en* (Engl. to consist in), *constar de* (Engl. to comprise), *servir para* (Engl. to serve for) and *significar* (Engl. to signify); around 120 for *concebir* (Engl. to conceive), *usar como / para* (Engl. to use as / for), *utilizar como / para* (Engl. to utilize as / for); and around 20 for *consistir de* (Engl. to consist of), *denominar también* (Engl. also denominated) and *llamar también* (Engl. also called).

The following restrictions were imposed on the search for verbal patterns:

- **Verbal form**: infinitive, participle and conjugate forms.
- **Verbal tense**: present and past for the simple forms, any verbal tense for the compound forms.
- **Person**: 3rd singular and plural for the simple forms, any for the compound forms.

The obtained occurrences were automatically annotated with contextual tags. These simple tags will act as delimiters during the next automatic process. For each occurrence, the definitional verbal pattern was annotated with "<dvp></dvp>"; everything before the pattern with "<left></left>"; everything after the pattern with "<right></right>"; and finally, in those cases where the verbal pattern includes a nexus, like the adverb *como* (as), everything between the verbal pattern and the nexus was annotated with <nexus></nexus>.

Here is an example of a DC with contextual tags:

> <left>El metabolismo</left> <dvp>puede definir se</dvp> <nexus>en términos generales como</nexus> <right>la suma de todos los procesos químicos (y físicos) implicados.</right>
>
> Engl. <left>Metabolism</left> <dvp>could be defined</dvp> <nexus>in general terms as</nexus> <right>the sum of all the chemic (and physic) implied processes</right>

4.3 Filtering of non-relevant contexts

Once we had extracted and annotated the occurrences containing DVPs, the next process was the filtering of non-relevant contexts. This was done based on the fact that definitional

patterns are not used only in definitional sentences. In the case of DVPs, some verbs tend to have a higher metalinguistic meaning than others. That is the case of *definir* (Engl. to define) or *denominar* (Engl. to denominate), vs. *concebir* (Engl. to conceive) or *identificar* (Engl. to identify), where the last two could be used in a wider variety of sentences. Moreover, the verbs with a high metalinguistic meaning are not used only for defining terms.

In a previous work, Alarcón and Sierra (2006) carried out an analysis to determine the type of grammatical particles or syntactic sequences that could appear when a DVP is not used to define a term. Those particles and sequences were found in some specific positions, for example: some negation particles like *no* (Engl. not) or *tampoco* (Engl. neither) were found in the first position before or after the DVP; adverbs like *tan* (Engl. so), poco (Engl. few) as well as sequences like *poco más* (Engl. not more than) were found between the definitional verb and the nexus *como*; also, syntactic sequences like adjective + verb were found in the first position after the definitional verb.

In table 3 we present the rules we have implemented in a script to filter non-relevant contexts.

Table 3. Rules for filtering non-relevant contexts

Position	**Grammatical particle | sequence**
___DVP	no | en ningún caso (in no case) | tampoco (neither) **</left>**
	para (for) **</left>**
DVP___NEXUS	**<nexus>** conjugated verb
	no nexus **</nexus>**
	[así | ya] (thus | already) nexus **</nexus>**
	[Tan | tanto] (so | as much) .* nexus **</nexus>**
	[más | poco | poco más] (more | few | not more than) nexus **</nexus>**
	[gerund | que (that) (sign)] nexus **</nexus>**
	"," nexus **</nexus>**
	Personal conjugated verb nexus **</nexus>**
NEXUS___	**<right>** no
	<right> [antes | cuan | para | si (before | how | for | if)]
	<right> (se (impersonal pronoun)) personal conjugated verb
	<right> adjetive verb
	<right> adjetive sign

With these rules, the script can recognise contexts like the following examples:

Rule: **NO <left>**

<left>En segundo lugar, tras el tratamiento eficaz de los cambios patológicos en un órgano pueden surgir problemas inesperados en tejidos que previamente **no</left>** <dvp>se identificaron</dvp> <nexus>como</nexus> <right>implicados clínicamente, ya que los pacientes no sobreviven lo suficiente.</right>

<left>Secondly, after the efficient treatment of pathologic changes in an organ, unexpected problems could appear in tissues which were previously **not</left>** <dvp>identified</dvp> <nexus>as</nexus> <right>clinically implied, because the patients do not survive long enough.</right>

Rule: **<nexus> CONJUGATED VERB**

<left>Ciertamente esta observación tiene una mayor fuerza cuando el número de categorías </left> <dvp>definidas</dvp> **<nexus>es** pequeño como</nexus> <right>en nuestro análisis.</right>

<left>Certainly, this observation become stronger when the number of categories</left> <dvp>defined</dvp> **<nexus> is** small as</nexus> <right>in our analysis.</right>

4.4 Identifying constitutive elements

Once the non-relevant contexts were filtered, the next process was the identification of main terms, definitions and pragmatic patterns (when they occur).

In Spanish, DCs, and depending on each DVP, the terms and definitions can appear in some specific positions. For example, in DCs with the verb *definir* (Engl. to define), the term could appear to the left, nexus or right position (T *se define como* D; *se define* T *como* D; *se define como* T D), while in DCs with the verb *significar* (Engl. to signify), terms can only appear in left position (T *significa* D).

Therefore, in this module the automatic process is highly related to deciding the positions where the constitutive elements can occur. We decided to use a decision tree (Alarcón 2006) to solve this problem, i.e., to detect the probable positions of terms, definitions and pragmatic patterns by means of logic inferences. We established some regular expressions to represent each constitutive element (the sign ".*" means any word o group of words):

Term	=	delimiter (determiner) + name + adjective. {0,2} .* delimiter
Pragmatic pattern	=	delimiter (sign) (preposition | adverb) .* (sign) delimiter
Definition	=	delimiter determiner + name .* delimiter

As well as in the filtering process, the contextual tags have functioned as delimiters to demarcate decision tree's instructions. In addition, each regular expression could function as a delimiter.

At a first level, the branches of the tree correspond to the different positions in which constitutive elements can appear (left, nexus or right). At a second level, the branches correspond to the regular expressions of each DC element. The nodes (branches conjunctions) correspond to decisions taken from the attributes of each branch and are also horizontally related by *If* or *If Not* inferences, and vertically through *Then* inferences. Finally, the leaves correspond to the assigned position for a constitutive element.

Hence, figure 3 shows an example of the decision tree inferences to identify left constitutive elements. In this figure, **TRE** = term regular expression, **PPRE** = pragmatic pattern regular expression and **DRE** = definition regular expression.

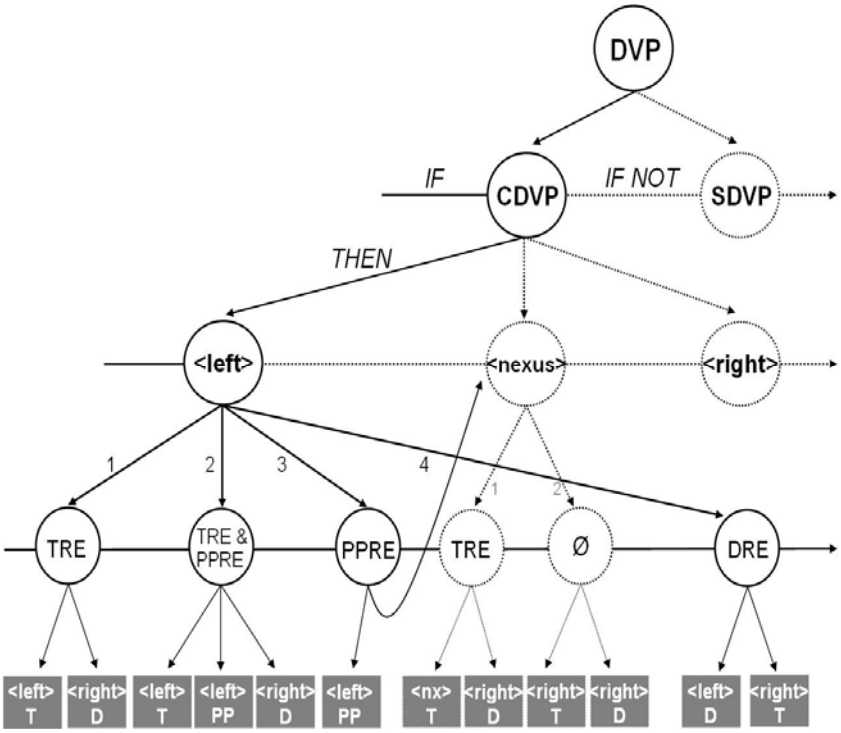


Figure 3. Identification of left position's constitutive elements

To illustrate this we can observe the following context:

→ <left>En sus comienzos</left> <dvp>se definió</dvp> <nexus>la psicología como </nexus> <right>"la descripción y la explicación de los estados de conciencia" (Ladd 1887).</right>

Once the DVP was identified as a CDVP –*definir como* (Engl. to define as)– the tree infers that left position:

1. Does not correspond only to a TRE.
2. Does not correspond to a TRE and a PPRE.
3. Does correspond only to a PPRE.

Then: left position is a pragmatic pattern (*En sus comienzos*), so to identify the term and its definition the tree goes to nexus inferences and finds that:

1. It does correspond only to a TRE.

Then: nexus position corresponds to the term (*la psicología*) and the right position corresponds to the definition ("la descripción y la explicación de los estados de conciencia […]").

The result consists of the processed context which was reorganised into terminological entries as shown in the example of table 4.

Table 4. Example of constitutive elements identification

Term	psicología
Definition	"la descripción y la explicación de los estados de la conciencia" (Ladd, 1887).
Verbal pattern	se define como
Pragmatic pattern	En sus comienzos

At this stage, the experiments helped us to define the best order to execute the inferences. The best results were obtained when the tree starts reading the nexus's position searching for regular expressions, continues to the left position and finalises searching at the right side.

Finally, it is important to mention that the distinction of genus and differentia in analytical definitions has not yet been implemented. We are in the process of developing scripts for this important task.

4.5 Evaluation

The evaluation of a system to extract definitional knowledge is not an easy task compared to the evaluation of other information extraction systems. A "gold standard" is quite difficult to establish, while the definition of a term could be more or less relevant depending on the specialised level of the evaluated texts as well as the evaluator criteria.

Trying to define a systematic way to evaluate our system, we decided to do it in two steps. We first evaluated the extraction of DVPs and the filtering of non–relevant contexts. Then, we evaluated the identification of the DCs elements. We describe each one in the next sections.

4.5.1 Evaluation of DVPs extraction and filtering of non-relevant contexts

We firstly evaluated the extraction of DVPs and the filtering of non-relevant contexts by means of Precision. Generally speaking, Precision measures how much automatically extracted information is *relevant*. To determine this, we used the following formula:

P = the total number of DCs automatically extracted, over the total number of contexts automatically extracted.

Precision was measured before and after the filtering process by analysing manually the results obtained from the DVP extraction and the filtering of non-relevant contexts. By the simple extraction of DVP occurrences we obtained the values shown in table 5 – Precision 1 column, whereas the values after the filtering process are shown in the Precision 2 column[2].

Table 5. Results of Precision1 and Precision2

Verbal pattern		**Precision 1**	**Precision 2**
concebir (como)	to conceive (as)	0,591	0,673
definir (como)	to define (as)	0,772	0,849
entender (como)	to understand (as)	0,287	0,342
identificar (como)	to identify (as)	0,256	0,311
consistir de	to consist of	0,588	0,625
consistir en	to consist in	0,592	0,601
constar de	to comprise	0,944	0,947
denominar también	also denominated	1	1
llamar también	also called	0,909	0,909
servir para	to serve for	0,528	0,556
significar	to signify	0,256	0,291
usar como	to use as	0,380	0,41
usar para	to use for	0,664	0,674
utilizar como	to utilise as	0,424	0,453
utilizar para	to utilise for	0,528	0,532

The average score for Precision in both cases was 0.60. Furthermore, after the filtering process these values slightly improved. We also noticed that there was a divergence on verbs usually appearing in metalinguistic sentences. The best results were obtained with verbs like *denominar* (Engl. to denominate) or *definir* (Engl. to define), while verbs like *entender* (Engl. to understand) or *significar* (Engl. to signify) had low Precision values. Verbs with lower results can be used in a wide assortment of sentences, (i.e., not necessarily definitional contexts), and they tend to recover a large quantity of noise.

The challenge we face at this stage is directly related to the elimination of noise. We have noticed that the more precise the verbal pattern is, the better results (in terms of less noise) can be obtained. However, the specification of verbal patterns would probably mean a reduced range of coverage. A revision of the filtering rules must be done in order to improve the identification of non-relevant contexts to avoid the cases when some DCs were incorrectly filtered.

The closest previous work to compare our results with is Malaisé et al. (2005), who report an average of up to 55% Precision which is fairly similar to the 60% we obtained.

4.5.2 Evaluation of DC's elements identification

We then evaluated the identification of DC's elements from the contexts filtered as DCs. To achieve this we assigned manually the following values to each DC processed by the decision tree:

3 for those contexts where the constitutive elements were correctly classified;

2 for those contexts where the constitutive elements were correctly classified, but some extra information was also classified (mainly extra words or punctuation marks in term position);

1 for those contexts where the constitutive elements were *not* correctly classified, (for example when terms were classified as definitions or vice versa).

Finally, the symbol **Ø** means the contexts that the system could not classify.

Table 6 shows the evaluation results for the identification of DC's elements. The values are expressed as percentages, and the amount of all of them represents the total number of DCs found with each verbal pattern.

Table 6. Evaluation of DCs elements identification

Verbal pattern		**3**	**2**	**1**	**Ø**
concebir (como)	to conceive (as)	68.57	15.71	11.42	04.28
definir (como)	to define (as)	65.10	18.22	10.41	06.25
entender (como)	to understand (as)	54.16	20.83	08.33	16.66
identificar (como)	to identify (as)	51.72	05.17	34.48	08.62
consistir de	to consist of	60.00	0	20.00	20.00
consistir en	to consist in	60.81	8.10	15.54	15.54
constar de	to comprise	58.29	22.97	02.97	15.74
denominar también	also denominated	21.42	28.57	07.14	42.85
llamar también	also called	30.00	40.00	0	30.00
servir para	to serve for	53.78	27.27	0.007	18.18
significar	to signify	41.26	44.44	03.17	11.11
usar como	to use as	63.41	14.63	17.07	04.87
usar para	to use for	36.26	32.96	04.39	26.37
utilizar como	to utilise as	55.10	28.57	10.20	06.12
utilizar para	to utilise for	51.51	19.69	10.60	18.18

From this table we would like to emphasise the following facts:

- The average percentage of the correctly classified elements (group "3") is over 50 percent of the global classification. In these cases, the classified elements correspond exactly to a term or a definition.
- In a low percentage (group "2"), the classified elements include extra information or noise. Nevertheless, in these cases the elements were also correctly classified as in group "3".
- The incorrect classification of terms and definitions (group "1"), as well as the unclassified elements (group "Ø") correspond to a low percentage of the global classification.

- There is also a different distribution of values among the treated verbs, since the percentage of group 3 versus the percentage of group 2 and 1 differs for each verb. In most cases the percentage of group 3 is higher than the percentage of groups 2 or 1. Nevertheless, in three cases the percentages of group 2 were higher than those of group 3.

Since the purpose of this process was the identification of DC's elements, and the average value obtained was over the 50% of corrected classified elements, we can argue that the results were generally satisfactory. However, there is still a lot of work to be done in order to improve the performance of the decision tree inferences. This work is related to the way the tree analyses the different DC's elements of each verbal pattern. At the moment, we have developed general inferences, but we recognise that particular inferences for those verbs with low recognition percentages need to improve to achieve the correct classification.

5. Conclusions

In this paper we have described the role of definitional verbal patterns to extract definitional knowledge. We have presented a set of semantic relations that link a definition with specific verbal patterns in a definitional context. This analysis was an important aim in order to design an automatic system for definitional knowledge extraction. This system, according to the test and the preliminary results we have obtained, is a relevant tool that could be helpful in the extraction of semantic relations in Spanish.

We are currently working on improving the rules for the filtering of non-relevant contexts process to perform a better identification of DCs, as well as improving the algorithm for the automatic process of identification of constitutive elements.

Although we have worked with definitional verbs, there is still a lot of work to be done in order to improve the system we have presented. We are currently working on the optimisation of the filtering rules to perform a better identification of DCs. It is necessary to continue with the formal description of all linguistic and metalinguistic patterns that constitute a DC and to observe the possible role that these other patterns play for establishing alternative semantic relationships between definitions.

Finally, we also have to explore other kinds of definitional patterns (mainly typographical patterns and reformulation markers) that are capable to recover definitional contexts.

Aknowledgments

This research was made possible by the financial support of the Consejo Nacional de Ciencia y Tecnología, Mexico (46832). The authors wish to thank the anonymous reviewers for its comments and suggestions, which helped to improve this paper, as well as Maria Pozzi for the proofreading of this paper.

Notes

[1] BwanaNet can be found at: http://bwananet.iula.upf.edu/indexes.htm
[2] A number close to 1 indicates a better result.

References

Alarcón, R. 2006. *Extracción automática de contextos definitorios en corpus especializados. Propuesta para el desarrollo de un ECCODE (extractor de candidatos a contextos definitorios).* Ph.D. Project Thesis. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.

Alarcón, R. and G. Sierra. 2003. "The role of verbal predications for definitional contexts extraction." In *Terminologie et Intelligence Artificielle (TIA 2003).* 11-20. Université de Strasbourg, Strasbourg, France.

Alarcón, R. and G. Sierra. 2006. "Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios." In Hernández, A. And J.L. Zechinelli (eds.). *Avances en la Ciencia de la Computación, VII Encuentro Nacional de Ciencias de la Computación*. 242-247. San Luís Potosí: MSCC.

Alshawi, H. 1987. "Processing dictionary definitions with phrasal pattern hierarchies." *Computational Linguistics* 13(3-4): 195-202.

Amsler, R. 1981. "A taxonomy for English nouns and verbs." In *Proceedings 19th Annual Meeting of the Association for Computational Linguistics.* 133-38. Stanford University, California.

Bach, C. 2005. "Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado." *Debate Terminológico*. Electronic Journal 1. http://www.riterm.net/revista/n_1/bach.pdf

Boguraev, B. and J. Pustejovsky. 1996. "Issues in text-based lexicon acquisition." In Boguraev, B. and J. Pustejovsky (eds.). *Corpus Processing for Lexical Acquisition.*

3-17. Cambridge, Mass.: MIT Press.
Bourigault, D., C. Fabre, C. Frérot, M.-P. Jacques and S. Ozdowska. 2005. "Syntex, analyseur syntaxique de corpus." In *Actes des 12èmes journées sur le traitement automatique des langues naturelles (TALN).* 2: 17-20. Dourdan, France.
Bourigault, D., I. Gonzalez–Mullier and C. Gros. 1996. "LEXTER, a natural language tool for terminology extraction." In *Proceedings of the Seventh EURALEX International Congress.* 771-779. Göteborg, Sweden.
Bowers, J. 2003. "Predication." In Baltin, M. and C. Collins (eds.). *The Handbook of Contemporary Syntactic Theory.* 299-333. Oxford: Blackwell.
Calzolari, N. and E. Picchi. 1988. "Acquisition of semantic information from an on-Line dictionary." In *Proceedings COLING-88.* 87-92. Budapest, Hungary.
Condamines, A. and J. Rebeyrolle. 2001. "Searching for and identifying conceptual relasionships via a corpus-based approach to a terminological knowledge base (CTKB)." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds.). *Recent Advances in Computational Terminology.* 127-148. Amsterdam/Philadelphia: John Benjamins.
Cruse, D. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
Chomsky, N. 1981. *Lectures on Government and Binding*. The Hague: Mouton de Gruyter.
Demonte, V. 1987. "C-command, prepositions and predication." *Linguistic Inquiry* 18(1): 147-157.
Feliu, J. 2004. *Relaciones conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Ph.D. Thesis. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.
Feliu, J., J. Vivaldi and M.T. Cabré. 2006. "SKELETON: Specialised knowledge retrieval on the basis of terms and conceptual relations." In *5th International Conference on Language Resources and Evaluation (LREC2006)*. LREC06, 2377-2382. Genoa, Italy.
Karttunen, L. and A. Zwicky. 1985. "Introduction." In Dowty, D., L. Karttunen and A. Zwicky (eds.). *Natural Language Parsing.* 1-25. Cambridge: Cambridge University Press.
Klavans, J. and S. Muresan. 2001. "Evaluation of the DEFINDER system for fully automatic glossary construction." In *Proceedings of the American Medical Informatics Association Symposium.* 252-262. New York: ACM Press.
Malaisé, V. 2005. *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels*. Ph.D. Thesis. Paris: Université Paris 7.
Malaisé, V., P. Zweigenbaum and B. Bachimont. 2005. "Mining defining contexts to help structuring differential ontologies." In Ibekwe-SanJuan, F., A. Condamines, and M.T. Cabré (eds.). *Terminology: Application-Driven Terminology Engineering.* 21-53. Amsterdam/Philadelphia: John Benjamins.
Mallén, E. 1991. "A Syntactic analysis of secondary predication in Spanish." *Journal of Linguistics* 27: 375-403.
Meyer, I. 2001. "Extracting knowledge-rich contexts for terminography." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds.). *Recent Advances in Computational Terminology*. 127-148. Amsterdam/Philadelphia: John Benjamins.

Morin, E. 1998. “Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes.” In *Actes, 5ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*. 172 –181. Paris.

Napoli, D. 1989. *Predication Theory*, Cambridge: Cambridge University Press.

Pearson, J. 1998. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

Pustejovsky, J., S. Bergler and P. Anick. 1993. “Lexical semantic techniques for corpus analysis.” *Computational Linguistics* 19(2): 331-58.

Rodríguez, C. 2004. “Metalinguistic information extraction for terminology.” In Ananiadou, S. and P. Zweigenbaum (eds.). *3rd International Workshop on Computational Terminology Coling-04.* 15-22. Geneva, Switzerland.

Sager, J. C. and A. Ndi-Kimbi. 1995. “The conceptual structure of terminological definition and their linguistic realisations: A report on research in progress.” *Terminology* 2(1): 61-85.

Saggion, H. 2004. “Identifying definitions in text collections for question answering.” In *Proceedings of the 4th International Conference on Language Resources and Evaluation.* 1927-1930. Lisbon, Portugal.

Saggion, H. and R. Gaizauskas. 2004 “Mining on-line sources for definition knowledge.” In *Proceedings of the Florida Artificial Intelligence Research Society 2004*. 61-66. Miami: Florida.

Sells, P. 1985. *Lectures on Contemporary Syntactic Theories.* Stanford: CSLI, Stanford University.

Sierra, G., R. Alarcón, A. Medina, A. and C. Aguilar. 2003. “Definitional contexts extraction from specialised texts.” In Lewandowska-Tomaszczyk, B. (ed.). *PALC 2003 Proceedings: Language, Corpora and E-Learning*, 21-31. Frankfurt: Peter Lang Publisher.

Storrer, A. and S. Wellinghoff. 2006. “Automated detection and annotation of term definitions in German text corpora.” In *5th International Conference on Language Resources and Evaluation (LREC2006)*. 275-295. Genoa, Italy.

Vossen, P. and A. Copestake. 1993. “Defaults in lexical representation.” In Briscoe, T., V. Paiva and A. Copestake (eds.). *Inheritance, Defaults and the Lexicon*. 246-274. Cambridge: Cambridge University Press.

Williams, E. 1980. “Predication.” *Linguistic Inquiry* 11(1): 203-238.

Authors addresses

Gerardo Sierra

Grupo de Ingeniería Lingüística, Instituto de Ingeniería

Universidad Nacional Autónoma de México

Mexico, DF

Mexico

gsierram@ii.unam.mx

Rodrigo Alarcón
Grupo de Ingeniería Lingüística, Instituto de Ingeniería
Universidad Nacional Autónoma de México
Mexico, DF
Mexico
ralarconm@ii.unam.mx

César Aguilar
Grupo de Ingeniería Lingüística, Instituto de Ingeniería
Universidad Nacional Autónoma de México
Mexico, DF
Mexico
caguilar@ii.unam.mx

Carme Bach
Grupo IULATERM
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona
Spain
carme.bach@upf.edu

About the authors

Gerardo Sierra is a full-time researcher of the Engineering Institute of National Autonomous University of Mexico (UNAM). He is the Leader of the Language Engineering Group, founded in 2000. He has received his PhD in Computational Linguistics at UMIST, Manchester, in 1999. He has taught courses of Introduction to Language Engineering, Corpus Linguistics and Text Mining at the UNAM and other universities. His research interests are in the area of Language Engineering, specifically

Computational Lexicography, Terminotics, Corpus Linguistics, Semantic Relations and Text Mining. He has published some important papers oriented to Language Engineering in the *International Journal of Lexicography*, *Terminology* and *Lecture Notes in Computer Science*.

Rodrigo Alarcón is a PhD Student in Language Sciences at the Institut Universitari de Lingüística Aplicada of UPF, Barcelona, since 2003. He has received a Diploma in Advanced Studies in Applied Linguistics in 2006. His PhD Research Dissertation focuses on developing a definitional contexts extraction system from specialised corpora in Spanish. His research interests are Corpus Linguistics, Information Extraction, Text Mining, and Language Engineering.

César Aguilar is a PhD in Linguistics student in the Department of Linguistics of UNAM, Mexico City, since 2003. He has received a Masters Degree on Hispanic Linguistic from the UNAM in 2003. His PhD Research Dissertation focuses in the description and analysis of definitional verbal patterns associated with definitions in DCs in Spanish. His research interests are Computational Lexicology, Formal Grammars for Automatic Parsing and Language Engineering.

Carme Bach is a professor of Catalan Linguistics at University Pompeu Fabra and a researcher of the consolidated research group IULATERM at Pompeu Fabra University (Barcelona, Spain). Her research focuses on General and Specialised Discourse Analysis, Lexicography and Corpus Linguistics. Her publications deal with connectives, reformulation and its importance in the process of specialised discourse construction and in extraction of semantic information of specialised discourse. Her PhD thesis, which is about reformulation markers, includes a lexicographical implementation prototype for these units.